

MontCAS

(Montana Comprehensive Assessment System)

English Language Proficiency Assessment

MontCAS English Language Proficiency (ELP) Assessment

Technical Report

2006-2007

Copyright © 2009, held by the Montana Office of Public Instruction. All rights reserved.

Printed in the U.S.A

Table of Contents

	Page
1. Purpose of the Technical Report	1
2. Scope of Work – Year 1	1
3. Description of the MontCAS ELP	1
3.1 Purpose of the MontCAS ELP	1
3.2 Structure of the MontCAS ELP	2
4. MontCAS ELP 2006-2007 Operational Form Construction	3
4.1 Structure of MontCAS ELP 2006-2007	4
4.2 Alignment of the MontCAS ELP	5
5. MontCAS ELP 2006-2007 Administration	6
5.1 Testing Window	6
5.2 Assessment Training	6
5.3 Examiner Scripts	6
5.4 Listening Test Administration	6
5.5 Setting for the Test	7
5.6 Timing	7
5.7 Prompting and Repeating Test Information	7
5.8 Testing Absentees	8
5.9 Testing Accommodations	8
6. MontCAS ELP 2006-2007 Test Security	9
6.1 Bar-Coding and Return of Secure Materials	9
6.2 Storage and Shredding of Secure Materials	9
7. MontCAS ELP 2006-2007 Scoring and Reporting	10
7.1 Scoring of Multiple-Choice Items	10
7.2 Writing Checklist	10
7.3 Scoring of Constructed-Response Items	10
7.4 Reporting	12

8. MontCAS ELP 2006-2007 Student Demographic Summary	14
8.1 Ethnicity of the Test Population	14
9. MontCAS ELP 2006-2007 Item Analyses	14
10. Scaling and Equating of the MontCAS ELP	17
11. Setting Standards on the MontCAS ELP	18
12. Reliability of the MontCAS ELP	20
13. Validity of the MontCAS ELP	25
13.1 Content-related Validity	25
13.2 Construct and Criterion-related Validity	25
14. References	27
15. Appendices	
1. MontCAS ELP Item Difficulty and Discrimination Data	28
2. MontCAS ELP Standards Setting Report	45
3. Mountain West Assessment Consortium Foundation Document: Introduction	63
4. Secure Materials Check-In Process	66
5. Score Reports Interpretation Guide	68

MontCAS English Language Proficiency (ELP) 2006-2007 Technical Report

1. Purpose of the Technical Report

The purpose of this report is to provide the Montana Office of Public Instruction (OPI) as well as Montana educators, citizens, researchers, and other interested parties with technical documentation for the development, administration, and reporting of the Fall 2006 Administration of the MontCAS English Language Proficiency Assessment (MontCAS ELP). This report includes evidence of the reliability and validity of the assessment as well as information on the appropriate use and interpretation of test scores.

2. Scope of Work – Year 1

This report covers the activities of Year 1 of the Contract between the State of Montana Office of Public Instruction and Questar Assessment, Inc. Year 1, which began on September 11, 2006 and ended September 10, 2007, included the following general activities: design, development, and distribution of the MontCAS ELP 2006-2007 test forms which were administered during Fall 2006; scoring tests, setting performance standards, and reporting test results.

3. Description of the MontCAS ELP

3.1 Purpose of the MontCAS ELP. The Montana English Language Proficiency Assessment (MontCAS ELP) is an assessment of English language proficiency for grades K-12. It is a modified version of an assessment developed for the Mountain West Consortium and designed to fulfill the requirements of ‘No Child Left Behind’ (NCLB) legislation. The MontCAS ELP assesses English proficiency in Listening, Speaking, Reading, and Writing, and reports scores in each of those language domains as well as in Comprehension (a combination of select items from the Listening and Reading test) and a total score, representing overall English proficiency. The MontCAS ELP was designed to assess the status of a student’s proficiency in English and to measure progress in attaining English proficiency.

The MontCAS ELP was designed to be administered to all students who have been identified as ‘limited English proficient’ (LEP) in the State of Montana. The process for identifying students as LEP is controlled at the district level and may include administering the Home Language Survey as well as one or more of a number of assessments. The instructions printed in the MontCAS ELP Examiner Manuals read as follows:

“Montana observes the federal definition of limited English proficiency. Both language impact and academic achievement must be considered when identifying LEP students. A student must be identified as one of the following:

1. an individual who was not born in the U.S. or whose native language is a language other than English;
2. an individual who comes from an environment where a language other than English is dominant;
3. an individual who is American Indian or Alaskan Native and who comes from an environment where a language other than English has had a significant impact on the individual’s level of English language proficiency.

The student must also have sufficient difficulty speaking, reading, writing, or understanding the English language to deny such an individual the opportunity to learn successfully in classrooms where the language of instruction is English or to participate fully in our society.”

The LEP population in the state of Montana is different from that of many other states. In Montana, up to 80% of the students identified as LEP are of American Indian descent and are very likely growing up in a community where English is the primary language. The English used in that community may very well be a nonstandard version. The uniqueness of student populations in the Western United States, including the prevalence of students of American Indians descent, was part of the impetus for the formation of the Mountain West Consortium. And the test development procedures (Matthews, 2007) took the characteristics of the student population in member states into consideration. Although the population in Montana includes a higher percentage students of American Indian descent, that population is not qualitatively different from that of other Mountain West member states.

3.2 Structure of the MontCAS ELP. The MontCAS ELP test forms are letter-coded to correspond to five grade/grade spans, as follows:

Grade Span	Forms
K	A
1-2	B1, B2
3-5	C1, C2
6-8	D1, D2
9-12	E1, E2

Within each grade span (other than K), there are two forms: Level 1 (i.e., B1, C1, D1, and E1) and Level 2 (i.e., B2, C2, D2, and E2). Level 1 forms are intended for LEP students with beginning or novice skills in English. So, it is appropriate for students in their first year in a U.S. school and possibly other LEP students who are not reading simple stories and writing simple sentences. All other students,

including those students who have more than basic English language skills, take the Level 2 (Intermediate) test.

Each test form—whether it is a Level 1 form or a Level 2 form—is divided into four subtests: Listening, Speaking, Reading, and Writing. Reading, Writing, and Listening are designed to be group administered (except to Kindergarten students) and may be administered in separate or consecutive testing sessions. The Speaking test is individually administered to all grade spans. Each LEP student is expected to be tested in all four areas, regardless of proficiency, and students must be tested with the test that corresponds to their grade in school. No off-grade-level testing is permitted. In addition, all four subtests administered to a student must be from the same level (1 or 2) form within a grade span. Only one test—the Kindergarten Reading test—has provisions for halting test administration based on a frustration-level rule.

The MontCAS ELP is a paper-and-pencil test. On the Kindergarten form, students either respond orally or circle their response in the test booklet. The examiner marks the answer document based on the student's response. On the Grade Span 1-2 forms, students administered the Level 2 assessment mark bubbles in their machine-scorable test booklet. In all other grade spans, Level 2 students mark or write their responses in a separate, scannable answer document. Note that Level 1 test booklet and answer documents (B1, C1, D1, and E1) were non-scannable to allow a cost saving due to the low quantity of tests needed and to allow for the limited time between the completion of Level 1 registration (to obtain Level 1 student counts) and materials distribution to systems.

4. MontCAS ELP 2006-2007 Operational Forms Construction

Forms administered in Fall 2006—designated MontCAS ELP Fall 2006—were based on Mountain West Form I and were previously administered in Idaho as the Idaho English Language Proficiency Assessment (IELA). Prior to administration as the IELA, Mountain West test forms were reviewed and modified in several ways. The modifications fell into three areas:

- Directions for test administration. Some of the text intended to be read by the test administrator or by the test taker was modified to clarify directions.
- Rubrics for open-ended items. Some of the rubrics used to guide test administrators in scoring open-ended items were modified. The purpose of these modifications was to clarify rules for scoring and, in some cases, to add to the list of acceptable and unacceptable responses for each score point.
- Addition of linking items. In order to create a psychometric link between level 1 and 2 forms in each grade span, a sample of items (usually 5 each in Reading and Writing) was chosen from level 1 forms and these items were added to corresponding level 2 forms as common, linking items. Within a grade cluster, Listening and Speaking subtests on level 1 and level 2 forms were identical. Thus all Listening and Speaking items were eligible to be used as linking items.

All items appearing on the MontCAS ELP 2006 assessment were from the Mountain West item pool. The forms into which those items were configured were previously administered in Idaho as the IELA. The items on forms administered as the MontCAS ELP were identical to those administered previously in Idaho. The cover page and page headers of the forms were changed prior to administration as the MontCAS ELP.

4.1 Structure of MontCAS ELP 2006-2007. Table 1 shows, for each MontCAS ELP 2006-2007 test form, the grade span in which it was administered and the numbers of items by item type in each language domain as well as the number of points represented by those items. The items and points in the Comprehension column do not contribute to the Totals shown in the last two columns because all Comprehension items were part of the Listening or Reading tests.

All Listening and Reading items were eligible to be included on the Comprehension test. Those items that assessed a lower-level reading skill (e.g., letter identification, sound-symbol correspondence) were not included as comprehension. In addition, stand-alone vocabulary items were not included although vocabulary-in-context items were included. Two individuals with extensive experience in test development independently identified those items on the Listening and Reading subtests that assessed comprehension. On those occasions where they disagreed, a third person evaluated the item and broke the tie.

Table 1. Structure and Content of MontCAS ELP 2006-2007 Test Forms

Form	Grade Cluster	Item Type	Listen		Speak		Read		Write		Comp		Total	
			Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts
A	K	MC	9	9	-	-	23	23	-	-	16	16	32	32
		SA	13	13	10	10	13	13	-	-	13	13	36	36
		ER	-	-	4	12	-	-	-	-	-	-	4	12
		Total	22	22	14	22	36	36	22*	22*	29	29	94	102

B1	1-2	MC	22	22	-	-	15	15	-	-	31	31	37	37
		SA	-	-	10	10	-	-	11	11	-	-	21	21
		ER	-	-	4	12	-	-	2	4	-	-	6	16
		Total	22	22	14	22	15	15	13	15	31	31	64	74
B2		MC	22	22	-	-	20	20	-	-	39	39	42	42
		SA	-	-	10	10	-	-	10	10	-	-	20	20
		ER	-	-	4	12	-	-	3	10	-	-	7	22
		Total	22	22	14	22	20	20	13	20	39	39	69	84

Form	Grade Cluster	Item Type	Listen		Speak		Read		Write		Comp		Total	
			Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts
C1	3-5	MC	22	22	-	-	15	15	4	4	31	31	41	41
		SA	-	-	10	10	-	-	5	5	-	-	15	15
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	31	31	62	74
C2		MC	22	22	-	-	18	18	9	9	37	37	49	49
		SA	-	-	10	10	1	2	-	-	1	2	11	12
		ER	-	-	4	12	-	-	3	10	-	-	7	22
		Total	22	22	14	22	19	20	12	19	38	39	67	83

D1	6-8	MC	22	22	-	-	15	15	5	5	32	32	42	42
		SA	-	-	10	10	-	-	4	4	-	-	14	14
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	32	32	62	74
D2		MC	22	22	-	-	18	18	10	10	38	38	50	50
		SA	-	-	10	10	-	-	-	-	-	-	10	10
		ER	-	-	4	12	2	6	3	10	2	6	9	28
		Total	22	22	14	22	20	24	13	20	40	44	69	88

E1	9-12	MC	22	22	-	-	15	15	7	7	32	32	44	44
		SA	-	-	10	10	-	-	2	2	-	-	12	12
		ER	-	-	4	12	-	-	2	6	-	-	6	18
		Total	22	22	14	22	15	15	11	15	32	32	62	74
E2		MC	22	22	-	-	19	19	10	10	39	39	51	51
		SA	-	-	10	10	-	-	-	-	-	-	10	10
		ER	-	-	4	12	2	6	3	10	2	6	9	28
		Total	22	22	14	22	21	25	13	20	41	45	70	89

* Items on the Kindergarten Writing test are configured as a checklist completed by the examiner.
MC - Multiple Choice; SA - Short Answer; ER - Extended Response

4.2 Alignment of the MontCAS ELP. An alignment study of the MontCAS ELP to the Montana English Language Proficiency Standards has not yet been completed. In the development of the Mountain West Consortium Test (Matthews, 2007), the member states of the consortium developed a set of common English language development (ELD) standards. The MWAC ELD standards were used to guide item development for the Mountain West Test.

5. MontCAS ELP 2006-2007 Administration

5.1 Testing Window. The testing window for MontCAS ELP 2006-2007 was November 28 through December 19, 2006. All test materials were to be returned to Questar by January 10, 2007.

5.2 Assessment Training. To prepare systems for the administration of the Fall 2006 MontCAS ELP, a Training PowerPoint Presentation was created. A Training CD with this presentation was shipped to all systems with a known LEP population on November 3, 2006 and the presentation was posted to the Office of Public Instruction website, <http://opi.mt.gov/assessment/ELP.html>. A Training Handout, which showed each slide from the Training Presentation, was also provided.

Each System Test Coordinator was encouraged to read through these presentations prior to administration and to consider using the PowerPoint presentations to train test administrators. Test Coordinators were also encouraged to take a Training Self-Quiz, which was also provided on the Training CD and as a PDF file on the Office of Public Instruction website, <http://opi.mt.gov/assessment/ELP.html>. The self quiz included some key questions regarding the test administration and answers were provided at the end of the quiz. Performance data from the self quiz were not available for evaluation.

To prepare for testing, examiners were instructed (in the examiner manual) to:

- read the manual completely;
- ensure that they had adequate materials for all students who would be tested;
- notify students in advance of testing;
- print and bubble on answer sheets student data of all students to be tested; and
- secure a CD player (or computer with CD-ROM drive, sound card and speakers) for administering the Listening test, and check the CD and the sound quality.

5.3 Examiner Scripts. Specific step-by-step instructions and script were provided for each test form in an examiner manual specific to that particular form. Scoring guides were provided for all oral constructed responses. Such items occurred throughout the Kindergarten form, but only in the Speaking test at all other grade spans. Where appropriate, examples of full-credit and partial-credit responses were provided.

5.4 Listening Test Administration. The Listening test was administered with a CD recording. This ensured that all students heard the questions in the same voice and at the same pace. The recording included a tone after each question signaling the examiner to pause the CD while students responded. A printed Listening Script for each form was available in the case that a school may request one.

5.5 Setting for the Test. For the individually administered subtests, examiners were advised as follows: “The test setting should be a quiet one-to-one environment. The testing should take place where other students cannot hear or see the testing materials. The examiner should sit close enough to the student to point to questions and illustrations in the student’s test booklet during test administration.”

For the group-administered subtests, examiners were advised as follows: “The test setting for the group-administered sections is a quiet classroom. The students should have in front of them only their test booklet, answer document, and a No. 2 pencil.”

It was also suggested that “Examiners should place a “Testing: Do Not Disturb” sign on the door of the testing site.”

5.6 Timing. The MontCAS ELP is an untimed test and examiners were advised to allow students as much time as they needed to finish any given subtest.

5.7 Prompting and Repeating Test Information. The following rules regarding prompting or repeating information were printed in all examiner manuals:

Prompting is the provision of additional information to students during administration of the assessment. Prompting includes

- elaborating on questions,
- clarifying information provided in reading selections or any test question,
- pointing out specific information in the questions or graphics,
- providing cues that might normally be part of an instructional strategy, and/or
- suggesting strategies that a student may use to arrive at a correct response.

In general, prompting is **not** allowed in this test because it may give an unfair advantage to some students. However, in specific situations where partial or unclear responses are given, the following general prompts are appropriate.

To clarify the student’s response, the examiner may say,

I don’t understand what you said.

Can you tell me more?

If the student answers in another language, the examiner may say,

Can you say that in English?

The examiner may repeat directions, if necessary, but must do so before the child begins a response.

If there is a distraction or interruption, the selection or question may be repeated.

If a student asks for a question to be repeated, the examiner may repeat the question only once.

If the student still does not understand what is being asked, the examiner should score that question as though the student gave no response (*BL*).

The examiner must not modify directions in any way. To do so would provide an unfair advantage to one student or a group of students over others.

The examiner should allow approximately 15 seconds of wait time for a student to begin a response to a question. This gives the student time to gather his or her thoughts and to think carefully before responding in English. If a student has not responded after 15 seconds, the examiner should move on to the next item or task and score the item as “no response” (*BL*).

5.8 Testing Absentees. Examiners were advised to make every effort to see that all LEP students in the school were administered all sections of the MontCAS ELP. If a student was absent for a particular testing session, a make-up test was to be scheduled, as long as it was within the testing window.

5.9 Testing Accommodations. For visually impaired students, the MontCAS ELP 2006-2007 was available (by special order) in Braille and in Large-Print. No Braille forms or Large-Print forms were ordered before or subsequent to the October 1, 2006, deadline.

For students with an Individual Education Plan (IEP) or 504 Plan on file, detailed instructions on Standard and Nonstandard Accommodations were provided in each Examiner Manual. In the Guidelines for Standard and Nonstandard Test Accommodations it was noted that some of the accommodations were crossed out on the listing and NA was coded in the accommodations section of answer documents. These crossed-out accommodations were not appropriate for MontCAS ELP students. Examiners were instructed to only bubble accommodations IF the accommodation was made for a student with special needs.

Examiners were warned that such accommodations should be used only when absolutely necessary and only with students with an IEP or 504 Plan on file with specific accommodations indicated. If a student was tested with accommodations, the examiner was instructed to mark the appropriate bubble (box 9) on the answer sheet.

Certain accommodations would necessarily invalidate test scores. The following list of non-allowable accommodations was provided in the Training PowerPoint presentation and Training Handouts:

The following accommodations are NOT allowed:

- Test administration in a language other than English.
- Translation of the assessment into another language.
- Translation of the assessment into sign language.
- Use of dictionaries or other reference aids. This includes both monolingual and bilingual dictionaries.
- Accepting responses in a language other than English.

(If students respond in their native language, the examiner may ask them if they can “say that in English.” If they can’t, the response is scored as ‘0’.)

The use of any of these accommodations will invalidate test scores.

6. MontCAS ELP 2006-2007 Test Security

6.1 Bar-Coding and Return of Secure Materials. All secure materials (test booklets, prompt books, Listening test CDs, and examiner manuals) were individually bar-coded. These secure test materials were scanned upon packing and distributing to systems and then scanned again upon return to Questar to account for materials. Test Coordinators were instructed to return all test materials—used and unused—to Questar. More detailed information about this process is included in Appendix 4.

6.2 Storage and Shredding of Secure Materials. After scoring, all used test booklets and answer documents were stored in Questar’s secure warehouse facility in Brewster, NY. Used answer documents are stored according to their processing so that they can be retrieved quickly, if necessary. Access to these facilities is limited to Questar staff. Used student answer documents must be stored for 180 days, and then Questar will obtain written permission from the State Manager to recycle the materials using a secure method of destruction. Questar received permission from the Montana Office of Instruction in February 2009 to destroy the used 2006-2007 materials.

All unused and non-scannable secure 2006-2007 materials were stored for 180 days. Except for file copies, all unused secure 2006-2007 test materials (i.e., examiner manuals, prompt books, and non-scannable test booklets) were shredded upon written permission from OPI.

7. MontCAS ELP 2006-2007 Scoring and Reporting

7.1 Scoring of Multiple-Choice Items. Multiple choice items (which are bubbled on the student test booklet or answer document) were scored electronically. One (1) point was given for the correct answer bubbled. Zero (0) points were given for incorrect answer bubbled or multiple bubbles marked. If no item was bubbled (an omit), the response was scored as a 'blank'.

7.2 Writing Checklist. The Writing raw score for Form A (Kindergarten level) was calculated as follows: 1 point was allocated for each skill on the Writing Checklist that the student "does most of the time" or of which they "demonstrate mastery." Thus, the Writing Checklist generated a maximum raw score of 22 points.

7.3 Scoring of Constructed-Response Items. The MontCAS ELP includes constructed-response (CR) items (separated into short answer [SA] and extended response [ER] in Table 1] in Speaking and Writing as well as a few CR items in Reading. Speaking CR items were scored by the test administrator at the time of test administration. Scoring guides and examples of full and partial-credit items were included as part of the Examiner Manual. Speaking responses were not recorded and no attempts were made to assess the validity or reliability of the rating of Speaking items.

Writing and Reading CR items were scored at the Questar scoring center using a 1-point, 2-point, or 4-point scale. The table that follows shows the grade spans, forms, levels, and Domains where there are Reading/Writing CR items. A second independent read was provided for 20% of the Level 2 Writing CR items. Level 1 Writing CR items were rated by the Questar Scoring Directors without a rescore due to the low quantities and non-scannable test booklets/answer documents for each Level 1 form.

CONSTRUCTED-RESPONSE ITEMS		
Grade Span	Forms	Level and Domain
1-2	Form B	Level 1 & Level 2 Writing
3-5, 6-8, 9-12	Forms C, D, E	Level 1 & Level 2 Writing; Level 2 Reading

Training Materials. A *Scoring Manual for Open-Ended Reading/Writing Responses* was used in the training of readers for scoring constructed-response items. A separate scoring manual was created for each grade span (B, C, D, and E). Questar's content specialists reviewed the scoring guides and rubrics for the constructed-response items, noted where there were weaknesses (if any) in the rubrics, and identified types of responses that will likely be seen in the operational responses. When necessary, sample responses were added to various items and score points to

present a more complete scoring guide (which consist of background information, the scoring rubrics, and annotated anchor responses) used to train readers.

Staffing. The scoring team consisted of two scoring directors and 16 readers. One director managed scoring of reading items and the other managed scoring of writing items. Initially, six readers were assigned to reading and ten readers to writing. When the readers assigned to reading items completed their scoring, they were retrained and joined the writing group. None of the readers were released during training or subsequent scoring due to poor performance. Readers were trained on each item by grade span prior to scoring any of the items in that grade span. Following the group training, the readers completed paired reads on individual items. As the scoring proceeded, Reader Reliability Statistics and Scorepoint Distribution Statistics were monitored for each reader on a daily basis.

Reader Reliability. The constructed-response items that were scored by two readers provide information on reader reliability. Data relevant to this issue are summarized in Table 2. This table shows, for each level 2 form for each item or set of items, the maximum point value of the item(s) (Pts), the number of student papers read twice (N), the percent of items on which the readers agreed exactly (% Exact), and the percent of items on which reader agreement was within +/-1 one score point (% Ex+Adj). All items, even those with maximum point values of 4, were at 100% exact + adjacent agreement.

Table 2. Summary of Reader Reliability for MontCAS ELP Constructed-response Items

Form	Domain	Item(s)	Pts	N	% Exact	% Ex + Adj
B2	W	1-5	1	402	95	100
		6-10	1	402	93	100
		11	2	402	80	100
		12-13	4	402	74	100
C2	W	10	2	586	77	100
		11-12	4	586	64	100
	R	19	2	586	91	100
D2	W	11	2	586	77	100
		12-13	4	586	66	100
	R	15	2	586	93	100
		20	4	586	75	100
E2	W	11	2	718	74	100
		12-13	4	718	62	100
	R	16	2	718	89	100
		21	4	718	74	100

Handscoring Issues. There were two issues that arose in the handscoring of the MontCAS ELP in 2006-07 which could be fixed by the administrators of the exam. There were instances where students wrote their responses outside of the designated response area and instances where students were administered one or more subtests from the wrong grade span. These errors can be avoided during administration by:

- Ensuring that the student is writing his or her response in the correct (designated) place so that, when scanned, it can be scored.
- Ensuring that each student has the correct test document for her/his grade and level.

Recommendations regarding these errors were incorporated into the training for the subsequent administrations of the MontCAS ELP.

7.4 Reporting. Student performance in each of the five language domains (Listening, Speaking, Reading, Writing, and Comprehension) was reported in terms of raw score, scaled score, and proficiency levels. Student performance was also reported on the overall (Total MontCAS ELP) test in terms of raw score, scaled score, and proficiency level. In February 2007, a panel of Montana educators met to set standards for the MontCAS ELP in the form of cut scores for each proficiency level by grade. Additional details of the standard setting process are included in Section 11 of this report and in the Appendices. The reported scores were defined in the *2006-2007 MontCAS ELP Assessment Score Reports Interpretation Guide* as:

“Raw Scores. The raw score is the total number of correct answers on multiple-choice items plus the number of points earned on open-ended items. Raw scores on the MontCAS ELP can only be compared for the same domain and the same test form. For example, a Form B1 raw score cannot be compared to a Form B2 raw score.

Note: The Writing raw score for (Kindergarten level) Form A was calculated as follows: 1 point was allocated for each skill on the Writing Checklist that the student "does most of the time" or of which they "demonstrate mastery." Thus, the Writing Checklist generated a maximum raw score of 22 points.

Scaled Scores. Scaled scores are derived from raw scores and provide results for alternate forms (e.g., B1 and B2) on a common scale. MontCAS ELP scaled scores can be compared for the same domain and the same grade-span test (A, B, C, D or E). For example, all Form C Reading scaled scores can be compared, regardless of whether the student took the C1 or the C2 Reading test. However, Form C scaled scores cannot be compared to Form D scaled scores.

Total MontCAS ELP Proficiency Levels. For the total score, four proficiency levels are reported: Novice (N), Nearing Proficiency (NP), Proficient (P), and Advanced (A). These are based on the total scaled score and provide a holistic estimate of the student's English

proficiency. It is important to note that students at the same overall Proficiency Level may have different profiles of competence across the language domains.

Domain Proficiency Levels. Within each domain, two proficiency levels are reported, based on the student's scaled score: Below Proficient (BP) and Proficient or Above (PA). (Individual language domain tests are not long enough to reliably provide more than two levels of proficiency.)”

Procedures for establishing Overall and Domain proficiency levels are detailed in section 11 of this report.

Incomplete Testing. Students were required to take all four language domain tests. If a student did not take one or more of the domain tests, the reports showed dashes in place of scores for that domain. The reported Total MontCAS ELP score was based on the domain tests for which there are scores. Thus, if a student failed to take the Speaking Test for whatever reason, the Total MontCAS ELP score was based on a raw score of zero in Speaking. The reported Comprehension scores—which were based on a subset of Listening and Reading scores—was affected in the same way if the student failed to take either the Listening or Reading Test.

Reports Shipment. MontCAS ELP 2006-2007 results packages were shipped to systems on July 6, 2007. The system and each of its schools had separate results packets. Below are the reports that were in each packet. Additionally copies (two copies for each school and system) of the *2006-2007 MontCAS ELP Assessment Score Reports Interpretation Guide* (SRIG) were included in the shipment. The SRIG included a sample of each report type with information for understanding the report. The guide also included information for using the MontCAS ELP results. The SRIG was also posted on the OPI website, <http://opi.mt.gov/assessment/ELP.html>. The SRIG can be found in Appendix 5.

MontCAS ELP System Packet – 2006-2007

- Table of Contents
- System Summary Reports by grade
- Copy of each School Summary Report
- Copy of each School Roster

MontCAS ELP School – 2006-2007

- Table of Contents
- School Summary Reports by grade
- School Rosters
- Individual Student Reports
- Student Labels
- Parent Reports

8. MontCAS ELP 2006-2007 Student Demographic Summary

Identification of a LEP student's ethnicity was provided by system personnel during the testing window (the information was bubbled in on the student answer document).

8.1 Ethnicity of the Test Population. Table 3 provides a breakdown of the MontCAS ELP 2006-2007 test population by ethnicity.

Table 3. MontCAS ELP 2006-2007 Test Population By Ethnicity

Grade	N	% American Indian or Alaskan Native	% White	% Other *
K	551	83.7	7.6	8.7
1	572	75.4	16.6	8.0
2	525	69.3	18.5	12.2
3	527	74.6	15.6	9.9
4	504	71.0	18.3	10.7
5	450	73.8	13.8	12.4
6	462	72.7	17.1	10.2
7	490	71.6	16.9	11.4
8	515	75.3	13.8	10.9
9	578	86.2	3.1	10.7
10	441	85.5	3.4	10.7
11	420	85.7	2.9	11.4
12	344	87.2	2.9	9.9

*Other (Blank, Asian, Hispanic, Black or African American, and Native Hawaiian/Other Pacific Islander)

Information on students' home language was not available for this report.

9. MontCAS ELP 2006-2007 Item Analyses

This section provides classical item-level statistics for all items administered on MontCAS ELP 2006-2007 forms. The p-value is presented as an index of item difficulty and the point-biserial correlation is presented as an index of item discrimination.

P-Values. For multiple-choice items, the p-value statistic is defined as the proportion of students that answer an item correctly. For constructed-response items, the p-value is reported as the average number of points out of the maximum number of possible points for an item. P-values range from zero to one (1.0). A high p-value means that an item is easy; a low p-value means that an item is difficult. Generally, it is desirable for tests to include items that span a range of difficulty.

Point-biserial correlations. The point-biserial correlation for each item is an index of the association between the item score and the total-test score. It shows how well the item discriminates between low-ability and high-ability students, where ability is inferred from the overall test score. Point-biserial correlation coefficients range between -1.0 and +1.0. High positive values indicate that a high-ability student is more likely (than a student with lower ability) to answer an item correctly and low negative values indicate that a low-ability student is more likely (than a student with higher ability) to answer an item correctly.

Table 4 shows the average p-value and range and median point-biserial correlation coefficients and range by language domain and test form. These data are only shown for level 2 forms because the numbers of level 1 forms administered were low even when aggregated across grades within a grade span. This table shows that there are some differences in both range and average p-value across language domains. For example, the average p-value in both Reading and Writing is lower in Grades 6-8 and 9-12 than the average p-value in Speaking and Listening. There are also some differences in the range; the maximum p-value in Reading in K and 9-12 is lower than the maximum in Reading in the other grade clusters.

Tables with item difficulty and discrimination data by item are included as **Appendix 1**. The tables in Appendix 1 present information by grade cluster, form, language domain, and item type (MC or CR). Because so few students were administered level 1 forms, item analyses were completed for level 2 forms only. The tables show for each item on each level 2 form the number of students (N) who were administered the item, the p-value and point-biserial correlation. For MC items, the tables show the percent of students choosing each responses alternative and the percent left blank. For CR items, the tables show the percent of students earning each score point. Analyses of test level data, including raw score descriptive statistics and test reliability measures, are reported in Table 7.

Table 4. Summary of MontCAS ELP 2006-2007 Item Difficulty and Discrimination by Grade span and Language Domain

Grade Span	Form	Domain	N	Item p-value		Point Biserial	
				Avg	Range	Med	Range
K	A	L	551	0.54	0.11 - 0.88	0.31	0.21 - 0.51
		S	551	0.68	0.34 - 0.89	0.40	0.36 - 0.52
		R	551	0.40	0.12 - 0.71	0.45	0.20 - 0.62
		W	551	0.35	0.07 - 0.83	0.40	0.10 - 0.54
1-2	B2	L	993	0.74	0.37 - 0.97	0.33	0.19 - 0.48
		S	993	0.82	0.55 - 0.94	0.33	0.25 - 0.49
		R	993	0.69	0.38 - 0.92	0.40	0.19 - 0.53
		W	993	0.53	0.25 - 0.83	0.51	0.33 - 0.65
3-5	C2	L	1,462	0.76	0.44 - 0.94	0.35	0.20 - 0.45
		S	1,462	0.88	0.68 - 0.98	0.30	0.23 - 0.50
		R	1,462	0.67	0.45 - 0.92	0.42	0.22 - 0.48
		W	1,462	0.75	0.34 - 0.92	0.45	0.28 - 0.58
6-8	D2	L	1,448	0.81	0.57 - 0.94	0.33	0.16 - 0.42
		S	1,448	0.84	0.65 - 0.93	0.52	0.38 - 0.58
		R	1,448	0.64	0.36 - 0.76	0.39	0.22 - 0.53
		W	1,448	0.69	0.37 - 0.94	0.48	0.25 - 0.59
9-12	E2	L	1,774	0.81	0.49 - 0.94	0.39	0.24 - 0.47
		S	1,774	0.89	0.69 - 0.98	0.27	0.21 - 0.47
		R	1,774	0.68	0.32 - 0.93	0.40	0.21 - 0.51
		W	1,774	0.65	0.15 - 0.89	0.37	0.06 - 0.56

10. Scaling and Equating of the MontCAS ELP

Initial scaling and equating of the 2006-2007 MontCAS ELP forms were completed on those forms when they were administered in Spring 2006 as the Idaho English Language Proficiency Assessment. The decision was made to use the Idaho data for item calibration, scaling and equating because the population to whom the forms were administered in Idaho (approximately 22,900) was larger than the population to whom the test was administered in Montana (approximately 6,300). Although the LEP populations in Idaho and Montana are significantly different (approximately 85% of the LEP students in Idaho are of Hispanic origin whereas approximately 85% of the LEP students in Montana are of American Indian origin), concerns about the small size of the sample in Montana outweighed concerns about differences in the student populations.

The raw score to scale score conversion tables produced for the IELA were used to produce scores for the MontCAS ELP. A brief summary of the calibration, scaling and equating as completed in Idaho follows. Item calibration, scaling, and equating were done within the framework of Item Response Theory (IRT). The Rasch Model (Rasch, 1960) for dichotomous items and the Partial Credit Model (Masters, 1982) for polytomous items were used as the IELA IRT model. The software used to implement these models was WINSTEPS, version 3.57.1 (Linacre & Wright, 2005). Within each grade span, all items on both forms (e.g., C1 and C2) were concurrently calibrated. Within each grade cluster (except K), there were common items on level 1 and level 2 forms. All of the speaking and listening items appeared on both level 1 and level 2 forms and a minimum of five items each in Reading and Writing were common to level 1 and 2 forms. The concurrent calibration procedure placed all items from both forms on the same Rasch item difficulty scale, effectively equating level 1 and 2 forms. By using the Rasch item parameter estimates from the concurrent calibration for just those items that are in each form, separate raw score to Rasch ability (theta) conversion tables were produced for each form. A linear transformation of theta values in each grade cluster produced raw score to scaled score conversion tables for each form. In Idaho, the scale was created in such a way that one or two performance levels were set to particular values. Although the same scale was used in Montana, performance levels on the MontCAS ELP were established by Montana educators in the 2007 MontCAS ELP standards setting (see section 11). That panel set cut scores for each proficiency level and grade.

11. Setting Standards on the MontCAS ELP

A formal MontCAS ELP Standard Setting was undertaken by Questar Assessment in collaboration with the Montana OPI. The sessions were conducted over three days, February 28-March 2, 2007. The methods and results are described fully in a report that is included as Appendix 2.

Table 5a shows, for each form and grade, the range of MontCAS ELP scaled scores corresponding to each proficiency level as determined by the standard setting process.

Table 5a. Total MontCAS ELP Scaled Scores Corresponding to Proficiency Levels

Total MontCAS ELP Proficiency Levels					
Form	Grade	Novice	Nearing Proficiency	Proficient	Advanced
A	K	Below 363	363-395	396-424	At or above 425
B1 or B2	1	Below 345	345-373	374-420	At or above 421
	2	Below 373	373-407	408-465	At or above 466
C1 or C2	3	Below 361	361-383	384-416	At or above 417
	4	Below 374	374-396	397-429	At or above 430
	5	Below 387	387-406	407-453	At or above 454
D1 or D2	6	Below 367	367-388	389-412	At or above 413
	7	Below 367	367-391	392-419	At or above 420
	8	Below 370	370-391	392-436	At or above 437
E1 or E2	9	Below 370	370-392	393-420	At or above 421
	10	Below 373	373-395	396-423	At or above 424
	11	Below 376	376-399	400-434	At or above 435
	12	Below 376	376-399	400-434	At or above 435

Table 5b shows scale score ranges corresponding to proficiency levels in each of the language domains (Listening, Speaking, Reading, Writing) and Comprehension. In the case of language domain tests, two proficiency levels are reported. Individual language domain tests do not include enough items to reliably report more than those two levels of proficiency. The language domain cuts were established in the following way. Once the total MontCAS ELP scores were finalized, those cuts were expressed as scaled scores. The theta that corresponds to the “proficient” cut score in each grade was then expressed as a language domain scaled score using the same linear transform that was used to go from language domain thetas to scaled scores.

Table 5b. Language Domain MontCAS ELP Scaled Scores Corresponding to Proficiency Levels

Form	Grade	Language Domain Proficiency Levels	
		Below Proficient	Proficient and above
A	K	Below 98	98 and above
B1 or B2	1	Below 91	91 and above
	2	Below 103	103 and above
C1 or C2	3	Below 92	92 and above
	4	Below 99	99 and above
	5	Below 103	103 and above
D1 or D2	6	Below 95	95 and above
	7	Below 96	96 and above
	8	Below 96	96 and above
E1 or E2	9-	Below 96	96 and above
	10	Below 98	98 and above
	11	Below 100	100 and above
	12	Below 100	100 and above

A more complete report on the MontCAS ELP Standards Setting is included as **Appendix 2**.

Table 6 shows the percent of students in each overall proficiency category as defined by the cut scores in Table 5a.

Table 6. Total MontCAS ELP Proficiency Level by Grade in 2006-2007

Grade	Percent in each Proficiency Category			
	2006-2007			
	N	NP	P	A
K	23	39	31	7
1	8	22	57	13
2	7	12	65	16
3	5	14	58	24
4	4	15	58	23
5	7	17	68	9
6	3	19	58	19
7	6	20	58	17
8	8	21	66	5
9	3	25	67	5
10	2	26	66	5
11	5	32	62	2
12	4	32	61	3

N = novice; NP = nearing proficient; P = proficient; A = Advanced.

12. Reliability of the MontCAS ELP

Data bearing on the reliability of MontCAS ELP 2006-2007 Test Forms are shown in the panels of Table 7. This table shows for each form and each language domain (and comprehension and the total test) the number of students (N) who were administered the form, coefficient Alpha, a measure of internal-consistency reliability, the maximum raw score attainable, and the mean, standard deviation, and standard error of measurement (SEM) in both raw score and scale score units. Number of students represents the number for whom there was a valid test score and may vary across language domains in a grade to the extent that there were students who did not attempt one or more of the language domain tests. There is a total score for each student regardless of whether or not all language domain tests were attempted. Data are aggregated by grade for level 2 forms but by grade span for level 1 forms due to the small numbers of students administered the latter.

This table shows that there are some tests and domains where reliability is low (e.g., Speaking on form C2 in Grade 5). There is no consistent pattern, however. There were some individual cases of low reliability when the test was administered in Idaho. Once again, however, there was no consistent pattern. Reliability is good, however, on the total test which is the level at which classification decisions are made.

Table 7. Reliability, Raw Score and Scale Score Descriptive Statistics for MontCAS ELP Test Forms by Grade

Grade K				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
A	Listening	544	0.83	22	12.0	4.4	1.84	98.1	18.9	7.89
	Speaking	536	0.83	22	13.2	4.7	1.91	102.3	21.5	8.79
	Reading	534	0.94	36	15.0	9.0	2.22	86.8	26.0	6.41
	Writing	513	0.93	22	8.2	5.5	1.51	79.0	29.1	7.97
	Comprehen	545	0.83	29	12.8	5.1	2.10	95.9	16.5	6.85
	Total	551	0.95	102	46.8	19.0	4.24	382.1	34.7	7.77

Grades 1-2				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B1	Listening	100	0.86	22	15.5	3.9	1.47	101.1	16.7	6.35
	Speaking	92	0.91	22	13.2	5.3	1.60	100.8	20.4	6.17
	Reading	76	0.95	15	11.6	2.4	0.56	102.2	17.5	4.03
	Writing	76	0.95	15	10.7	3.6	0.78	107.2	24.7	5.36
	Comprehen	101	0.90	31	20.4	6.2	1.97	95.6	17.4	5.49
	Total	104	0.94	74	42.9	16.1	3.81	379.1	53.9	12.77

Grade 1				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B2	Listening	484	0.75	22	15.0	3.4	1.69	98.8	13.9	6.91
	Speaking	481	0.78	22	15.6	4.0	1.85	109.9	16.7	7.77
	Reading	485	0.78	20	11.7	3.9	1.83	93.1	16.9	8.00
	Writing	488	0.79	20	6.1	3.5	1.61	82.2	17.9	8.18
	Comprehen	489	0.82	39	24.3	6.1	2.59	95.4	13.2	5.58
	Total	489	0.88	84	47.8	11.4	3.89	389.0	29.5	10.09

Grade 2				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B2	Listening	495	0.80	22	18.2	2.4	1.08	113.8	14.1	6.35
	Speaking	486	0.84	22	17.7	3.3	1.31	119.0	17.0	6.80
	Reading	504	0.79	20	16.0	3.3	1.52	114.7	18.6	8.59
	Writing	502	0.81	20	11.9	4.1	1.77	111.1	23.2	10.11
	Comprehen	504	0.85	39	31.4	5.4	2.12	113.1	15.0	5.83
	Total	504	0.91	84	62.9	11.7	3.55	433.0	38.5	11.69

Grades 3-5				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C1	Listening	18	0.93	22	14.3	5.9	1.57	98.9	19.3	5.13
	Speaking	18	0.92	22	15.3	6.4	1.83	99.9	24.2	6.93
	Reading	14	0.95	15	7.9	5.3	1.15	91.0	27.2	5.90
	Writing	14	0.95	15	8.2	5.1	1.18	92.9	26.6	6.12
	Comprehen	18	0.92	31	17.0	7.3	2.06	94.6	18.0	5.08
	Total	19	0.96	74	40.0	19.0	3.88	381.0	36.3	7.44

Grade 3				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C2	Listening	514	0.82	22	15.5	4.0	1.67	101.5	13.2	5.55
	Speaking	506	0.83	22	17.6	3.5	1.46	106.7	15.0	6.27
	Reading	509	0.81	20	12.0	3.9	1.71	101.5	12.6	5.58
	Writing	507	0.79	19	11.1	3.3	1.52	102.0	15.0	6.81
	Comprehen	521	0.86	39	24.6	6.8	2.59	100.6	11.4	4.36
	Total	522	0.90	83	54.9	12.6	3.89	401.8	21.8	6.76

Grade 4				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C2	Listening	495	0.79	22	17.1	3.4	1.56	106.9	13.2	6.10
	Speaking	491	0.80	22	18.6	3.0	1.37	110.7	14.8	6.71
	Reading	494	0.83	20	13.9	4.1	1.71	107.6	14.4	6.00
	Writing	491	0.77	19	12.8	3.3	1.56	110.5	16.7	8.01
	Comprehen	498	0.86	39	28.1	6.4	2.43	106.3	12.1	4.56
	Total	499	0.91	83	61.6	11.8	3.62	413.5	23.0	7.05

Grade 5				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C2	Listening	439	0.78	22	18.0	3.2	1.48	111.4	13.7	6.37
	Speaking	441	0.68	22	19.3	2.9	1.63	114.3	15.2	8.63
	Reading	426	0.87	20	15.2	3.8	1.34	112.7	14.8	5.25
	Writing	437	0.76	19	13.8	3.2	1.58	116.4	18.9	9.35
	Comprehen	440	0.88	39	30.0	6.4	2.26	110.5	12.3	4.36
	Total	441	0.90	83	65.6	10.8	3.42	423.1	23.2	7.34

Grades 6-8				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D1	Listening	19	0.83	22	12.6	5.0	2.04	90.8	10.9	4.45
	Speaking	18	0.93	22	10.4	7.8	2.08	83.7	24.2	6.46
	Reading	19	0.72	15	7.8	3.2	1.71	86.0	11.8	6.22
	Writing	17	0.86	15	8.1	3.5	1.30	87.2	12.9	4.78
	Comprehen	19	0.82	32	16.9	6.1	2.55	88.9	8.4	3.53
	Total	19	0.94	74	37.4	16.1	4.11	370.1	21.7	5.52

Grade 6				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D2	Listening	452	0.79	22	17.1	3.6	1.63	100.7	9.8	4.47
	Speaking	451	0.76	22	18.5	3.1	1.49	104.5	11.5	5.61
	Reading	450	0.79	24	13.5	4.3	1.98	99.9	9.8	4.50
	Writing	448	0.78	20	11.7	3.6	1.67	99.5	11.2	5.23
	Comprehen	454	0.86	43	28.7	7.1	2.67	99.6	8.6	3.24
	Total	455	0.91	88	60.2	12.3	3.78	399.2	16.8	5.17

Grade 7				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D2	Listening	480	0.79	22	18.2	3.1	1.42	104.6	10.4	4.84
	Speaking	445	0.91	22	18.8	3.1	0.94	105.8	11.6	3.51
	Reading	466	0.83	24	14.9	4.1	1.66	103.1	9.5	3.89
	Writing	462	0.81	20	12.6	3.2	1.38	102.4	10.7	4.67
	Comprehen	484	0.86	44	30.7	6.9	2.61	102.3	9.0	3.42
	Total	484	0.93	88	61.8	14.1	3.87	402.1	19.4	5.32

Grade 8				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D2	Listening	495	0.89	22	18.7	3.4	1.12	106.8	11.6	3.78
	Speaking	461	0.91	22	18.2	3.4	1.01	103.4	11.7	3.49
	Reading	492	0.85	24	15.3	4.4	1.68	104.3	10.5	4.03
	Writing	486	0.82	20	12.9	3.5	1.46	103.3	11.9	5.01
	Comprehen	504	0.91	44	31.6	7.7	2.32	103.9	10.5	3.16
	Total	509	0.94	88	61.7	15.7	3.88	402.9	22.0	5.44

Grades 9-12				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E1	Listening	9	0.92	22	15.6	6.1	1.69	96.6	16.9	4.68
	Speaking	9	0.92	22	13.3	7.5	2.19	85.8	20.8	6.02
	Reading	9	0.90	15	9.8	4.5	1.44	94.2	17.7	5.59
	Writing	9	0.85	15	7.6	4.5	1.73	90.4	20.0	7.65
	Comprehen	9	0.94	32	21.6	9.0	2.12	95.9	17.3	4.10
	Total	9	0.97	74	46.2	20.6	3.79	386.2	25.0	4.61

Grade 9				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E2	Listening	564	0.83	22	17.6	3.2	1.30	99.6	10.3	4.24
	Speaking	563	0.81	22	18.3	3.3	1.42	101.9	10.5	4.57
	Reading	564	0.79	25	15.6	4.3	1.96	100.4	9.4	4.28
	Writing	563	0.73	20	11.2	3.3	1.70	100.4	9.9	5.15
	Comprehen	569	0.87	44	31.2	7.0	2.48	99.3	8.8	3.15
	Total	575	0.91	89	61.6	12.8	3.91	398.4	14.5	4.42

Grade 10				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E2	Listening	432	0.83	22	18.1	3.3	1.35	101.6	11.1	4.61
	Speaking	435	0.72	22	18.7	3.0	1.57	103.1	10.6	5.60
	Reading	431	0.80	25	16.6	4.3	1.93	102.7	9.9	4.48
	Writing	431	0.71	20	11.9	3.1	1.69	102.6	9.9	5.33
	Comprehen	435	0.87	45	32.7	6.9	2.51	101.6	9.5	3.44
	Total	437	0.90	89	64.7	11.5	3.74	402.2	14.3	4.63

Grade 11				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E2	Listening	410	0.87	22	18.5	3.4	1.20	103.6	12.3	4.38
	Speaking	412	0.81	22	18.8	3.0	1.35	103.9	11.1	4.91
	Reading	411	0.83	25	17.1	4.6	1.90	104.0	11.0	4.58
	Writing	412	0.73	20	12.4	3.2	1.69	104.1	10.3	5.38
	Comprehen	415	0.89	45	33.4	7.4	2.43	103.0	10.7	3.53
	Total	418	0.92	89	65.7	12.8	3.72	404.1	16.1	4.66

Grade 12				Raw Scores				Scale Scores		
Form	Language Domain	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E2	Listening	332	0.89	22	18.7	2.9	0.96	103.8	11.4	3.73
	Speaking	338	0.80	22	18.9	3.0	1.32	104.0	11.0	4.89
	Reading	336	0.83	25	17.3	4.3	1.78	104.1	10.4	4.32
	Writing	336	0.74	20	12.3	3.2	1.62	103.7	10.5	5.31
	Comprehen	339	0.90	45	33.6	7.0	2.24	102.9	9.8	3.13
	Total	344	0.92	89	65.5	12.9	3.75	403.5	15.6	4.55

13. Validity of the MontCAS ELP

13.1 Content-related Validity. Validity of the MontCAS ELP begins with test content. The Introduction to the Mountain West Assessment Consortium Foundation Document, included as **Appendix 3**, provides background information on the design of the assessment. Additional information on the development of the Mountain West items is provided in Matthews (2007).

13.2 Construct and Criterion-related Validity. In addition to test design considerations, test results also bear on the content validity of the assessment. In very general terms, the distribution and range of scores within each grade span and grade level (Table 7) provide evidence that the MontCAS ELP can capture a range of abilities. And, Table 8 provides information on the validity of the assessment showing intercorrelations among components of the test. This table shows, by grade span for level 2 forms, Pearson product moment correlations among scaled scores on each subtest (Listening, Speaking, Reading, Writing, and Comprehension). Correlations are not reported for subtests that share common items (e.g., Reading and Comprehension) nor are they reported for subtests and Total MontCAS ELP. The number below the correlation coefficient in each cell represents the number of students on which the correlation is based.

Table 8. Correlations Among Scaled Scores on Individual Language Domain Tests

Grade	K	1-2	3-5	6-8	9-12	
r	A	B2	C2	D2	E2	Avg.
L x S	0.68 535	0.39 964	0.35 1,427	0.25 1,345	0.27 1,716	0.39
L x R	0.46 533	0.61 975	0.48 1,418	0.58 1,393	0.58 1,722	0.54
L x W	0.30 506	0.62 976	0.49 1,425	0.54 1,383	0.48 1,724	0.49
S x R	0.41 529	0.36 963	0.32 1,406	0.28 1,347	0.33 1,721	0.34
S x W	0.26 501	0.38 964	0.33 1,416	0.25 1,339	0.24 1,721	0.29
S x C	0.63 536	0.40 967	0.37 1,435	0.30 1,351	0.34 1,732	0.41
R x W	0.38 500	0.74 986	0.63 1,418	0.66 1,392	0.63 1,735	0.61
W x C	0.34 507	0.74 990	0.65 1,434	0.68 1,396	0.62 1,742	0.61
Avg.	0.43	0.52	0.45	0.44	0.44	0.46

All of the correlation coefficients in Table 8 are significantly different from zero, indicating that the different subtests are measuring related abilities. Insofar as the language domain tests are measuring aspects of the same construct, English proficiency, performance in the different domains should be related. In addition, however, the coefficients are not high enough to suggest that the abilities measured by the individual domain tests are identical, reinforcing the assumption that language domain abilities are different aspects of overall English proficiency.

Additional evidence bearing on the validity of the MontCAS ELP (e.g., relation of test performance to that on other assessments or to classroom performance) was not available for this report.

References

- Linacre, J. M. & Wright, B. D. (2005). *A user's guide to WINSTEPS: Rasch-model computer program* (v. 3.57). Chicago, IL: MESA Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Matthews, G. (2007). Developing the Mountain West assessment. In J. Abedi (Ed.). *English language proficiency in the nation*, (pp. 33-45). Davis, CA: University of California, School of Education..
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Appendix 1: Item Difficulty and Discrimination data.

Grade K (Form A) Listening Items – MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
3	551	0.83	0.32	9.3	4.7	82.8		2.2
4	551	0.88	0.24	88.2	4.7	4.0		2.0
5	551	0.88	0.28	6.9	88.0	2.0		2.2
10	551	0.33	0.21	41.7	33.2	19.8		4.2
13	551	0.67	0.27	13.3	15.4	67.2		3.3
16	551	0.46	0.21	45.6	25.6	21.2		6.7
17	551	0.65	0.29	16.5	65.0	11.4		6.2

Grade K (Form A) Reading Items – MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	551	0.71	0.37	8.7	70.6	14.7		3.6
2	551	0.7	0.35	69.5	7.1	18.0		3.1
3	551	0.71	0.42	19.6	4.5	70.8		2.7
19	551	0.52	0.61	6.5	7.1	51.7		32.7
20	551	0.53	0.58	6.0	6.0	52.6		33.4
21	551	0.56	0.55	1.3	55.9	7.8		33.0
22	551	0.3	0.38	30.5	16.3	9.3		42.1
23	551	0.29	0.45	21.2	29.4	6.2		41.4
24	551	0.28	0.37	27.8	18.9	9.4		41.9
28	551	0.36	0.52	35.8	7.8	7.1		47.0
29	551	0.17	0.28	19.2	16.7	13.1		49.0
30	551	0.14	0.23	8.5	27.2	14.3		47.9
31	551	0.24	0.37	24.3	10.2	9.3		54.5
32	551	0.12	0.3	19.1	10.2	12.3		56.6
33	551	0.16	0.28	12.3	12.5	16.2		57.2
34	551	0.17	0.2	16.7	12.5	9.8		59.2
35	551	0.17	0.28	14.9	16.7	7.1		59.5
36	551	0.15	0.29	11.6	15.1	10.2		61.2

Grade K (Form A) Listening Items – CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	551	0.61	0.51	39.2	60.8	-	-	-
2	551	0.4	0.3	59.5	40.5	-	-	-
6	551	0.64	0.46	35.9	64.1	-	-	-
7	551	0.76	0.42	24.1	75.9	-	-	-
8	551	0.58	0.45	41.9	58.1	-	-	-
9	551	0.57	0.39	43.4	56.6	-	-	-
11	551	0.5	0.33	49.6	50.5	-	-	-
12	551	0.11	0.27	88.9	11.1	-	-	-
14	551	0.54	0.28	46.3	53.7	-	-	-
15	551	0.75	0.43	24.7	75.3	-	-	-
18	551	0.22	0.36	77.7	22.3	-	-	-
19	551	0.35	0.36	65.0	35.0	-	-	-
20	551	0.47	0.28	52.8	47.2	-	-	-
21	551	0.14	0.22	85.8	14.2	-	-	-
22	551	0.45	0.42	55.2	44.8	-	-	-

Grade K (Form A) Speaking Items – CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	551	0.87	0.38	12.9	87.1	-	-	-
2	551	0.88	0.38	12.2	87.8	-	-	-
3	551	0.56	0.4	43.7	56.3	-	-	-
4	551	0.86	0.4	14.0	86.0	-	-	-
5	551	0.83	0.4	16.5	83.5	-	-	-
6	551	0.77	0.38	23.4	76.6	-	-	-
7	551	0.89	0.36	11.3	88.8	-	-	-
8	551	0.56	0.39	44.5	55.5	-	-	-
9	551	0.76	0.48	24.0	76.0	-	-	-
10	551	0.79	0.42	21.2	78.8	-	-	-
11	551	0.45	0.51	38.3	33.6	28.1	-	-
12	551	0.59	0.48	17.4	47.0	35.6	-	-
13	551	0.41	0.52	27.6	19.4	25.4	17.4	10.2
14	551	0.34	0.49	29.2	29.0	24.1	13.4	4.2

Grade K (Form A) Reading Items – CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
4	551	0.66	0.44	34.3	65.7	-	-	-
5	551	0.55	0.45	44.8	55.2	-	-	-
6	551	0.62	0.42	37.9	62.1	-	-	-
7	551	0.57	0.45	42.7	57.4	-	-	-
8	551	0.54	0.54	46.3	53.7	-	-	-
9	551	0.41	0.51	59.4	40.7	-	-	-
10	551	0.46	0.62	53.7	46.3	-	-	-
11	551	0.42	0.59	57.9	42.1	-	-	-
12	551	0.47	0.62	53.2	46.8	-	-	-
13	551	0.65	0.68	35.4	64.6	-	-	-
14	551	0.52	0.6	48.3	51.7	-	-	-
15	551	0.63	0.65	37.2	62.8	-	-	-
16	551	0.39	0.57	60.8	39.2	-	-	-
17	551	0.45	0.61	55.0	45.0	-	-	-
18	551	0.17	0.34	83.1	16.9	-	-	-
25	551	0.28	0.53	71.7	28.3	-	-	-
26	551	0.27	0.56	73.0	27.0	-	-	-
27	551	0.23	0.51	77.0	23.1	-	-	-

Grade K (Form A) Writing Items – CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	551	0.67	0.1	32.7	67.3	-	-	-
2	551	0.66	0.48	33.9	66.1	-	-	-
3	551	0.57	0.42	42.8	57.2	-	-	-
4	551	0.62	0.45	38.3	61.7	-	-	-
5	551	0.83	0.31	17.2	82.8	-	-	-
6	551	0.35	0.45	64.6	35.4	-	-	-
7	551	0.41	0.33	59.0	41.0	-	-	-
8	551	0.21	0.39	78.8	21.2	-	-	-
9	551	0.42	0.54	58.3	41.7	-	-	-
10	551	0.26	0.48	73.5	26.5	-	-	-
11	551	0.16	0.42	84.2	15.8	-	-	-
12	551	0.57	0.45	42.7	57.4	-	-	-
13	551	0.58	0.46	41.9	58.1	-	-	-
14	551	0.38	0.46	62.1	37.9	-	-	-
15	551	0.15	0.4	84.6	15.4	-	-	-
16	551	0.15	0.37	84.6	15.4	-	-	-
17	551	0.09	0.32	91.1	8.9	-	-	-
18	551	0.12	0.38	88.0	12.0	-	-	-
19	551	0.07	0.25	93.5	6.5	-	-	-
20	551	0.21	0.32	78.8	21.2	-	-	-
21	551	0.09	0.31	91.3	8.7	-	-	-
22	551	0.08	0.31	91.7	8.4	-	-	-

Grade 1-2 (Form B-2) Listening Items – MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	993	0.94	0.29	94.1	0.2	1.2		0.7
2	993	0.91	0.34	3.9	0.5	90.5		3.4
3	993	0.97	0.25	96.6	1.1	0.3		0.5
4	993	0.93	0.28	93.5	1.6	2.3		1.0
5	993	0.83	0.37	12.1	1.9	83.3		0.9
6	993	0.92	0.31	92.3	1.2	3.3		1.8
7	993	0.97	0.37	1.1	96.7	0.4		0.3
8	993	0.59	0.31	8.1	59.3	28.0		1.4
9	993	0.82	0.31	9.5	82.2	6.7		0.1
10	993	0.65	0.44	16.3	65.4	10.7		5.7
11	993	0.76	0.42	6.9	10.9	76.3		3.4
12	993	0.82	0.38	81.7	10.6	3.3		2.6
13	993	0.83	0.42	83.2	3.7	7.6		2.3
14	993	0.62	0.42	22.4	62.1	9.2		4.2
15	993	0.71	0.48	70.9	7.9	14.4		5.2
16	993	0.84	0.41	83.7	7.1	4.4		1.6
17	993	0.89	0.41	5.1	1.4	88.9		2.2
18	993	0.37	0.27	35.5	37.4	17.1		8.5
19	993	0.59	0.26	29.0	3.6	59.4		5.5
20	993	0.44	0.27	24.1	44.3	21.5		8.8
21	993	0.57	0.19	57.4	23.3	10.7		5.0
22	993	0.39	0.29	26.3	23.1	39.5		9.6

Grade 1-2 (Form B-2) Reading Items – MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	993	0.79	0.33	10.6	78.6	6.1		4.0
2	993	0.52	0.43	52.2	27.6	13.3		6.5
3	993	0.84	0.46	9.5	4.5	83.9		1.6
4	993	0.82	0.44	6.0	8.8	82.5		2.2
5	993	0.77	0.48	7.5	12.4	77.1		2.3
6	993	0.84	0.41	84.5	6.5	4.1		4.2
7	993	0.92	0.32	3.9	91.9	2.1		1.2
8	993	0.8	0.32	80.4	13.1	3.3		2.7
9	993	0.53	0.4	53.4	11.4	32.0		1.9
10	993	0.89	0.46	5.1	89.1	2.6		2.7
11	993	0.77	0.5	9.1	10.6	76.7		3.0
12	993	0.56	0.45	14.5	55.6	25.4		3.3
13	993	0.7	0.5	17.1	69.6	7.0		5.9
14	993	0.72	0.37	11.5	71.7	10.5		5.8
15	993	0.48	0.19	33.2	12.7	48.2		5.1
16	993	0.53	0.29	19.1	53.2	16.3		11.0
17	993	0.69	0.53	10.5	9.3	68.8		11.1
18	993	0.38	0.25	38.3	28.6	20.8		11.9
19	993	0.87	0.34	5.0	86.7	1.1		6.3
20	993	0.42	0.36	20.4	23.4	41.9		13.9

Grade 1-2 (Form B-2) Speaking Items –CR

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	993	0.94	0.27	5.5	94.5	-	-	-
2	993	0.79	0.41	21.0	79.1	-	-	-
3	993	0.75	0.35	24.7	75.3	-	-	-
4	993	0.87	0.38	13.0	87.0	-	-	-
5	993	0.89	0.31	11.3	88.7	-	-	-
6	993	0.91	0.25	9.5	90.5	-	-	-
7	993	0.94	0.31	6.0	94.0	-	-	-
8	993	0.93	0.25	7.0	93.1	-	-	-
9	993	0.92	0.32	8.4	91.6	-	-	-
10	993	0.93	0.3	7.1	93.0	-	-	-
11	993	0.7	0.41	8.0	44.9	47.1	-	-
12	993	0.72	0.45	10.5	34.7	54.8	-	-
13	993	0.58	0.44	13.6	14.7	19.8	29.8	22.1
14	993	0.55	0.49	10.6	18.6	26.2	28.2	16.4

Grade 1-2 (Form B-2) Writing Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	993	0.66	0.49	33.9	66.1	-	-	-
2	993	0.39	0.54	61.0	39.0	-	-	-
3	993	0.71	0.53	28.7	71.3	-	-	-
4	993	0.69	0.46	31.2	68.8	-	-	-
5	993	0.83	0.4	16.6	83.4	-	-	-
6	993	0.45	0.33	54.8	45.2	-	-	-
7	993	0.67	0.51	33.3	66.7	-	-	-
8	993	0.38	0.44	62.1	37.9	-	-	-
9	993	0.37	0.46	62.6	37.4	-	-	-
10	993	0.52	0.58	48.1	51.9	-	-	-
11	993	0.65	0.58	13.4	43.4	43.2	-	-
12	993	0.26	0.65	34.5	36.1	21.1	6.7	1.7
13	993	0.25	0.59	40.5	30.7	19.9	7.0	1.9

Grade 3-5 (Form C-2) Listening Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1462	0.84	0.38	83.8	5.5	4.7	3.5	1.0
2	1462	0.89	0.3	1.7	4.2	1.2	88.6	1.4
3	1462	0.93	0.33	1.1	93.1	1.2	2.3	1.0
4	1462	0.87	0.37	3.1	87.4	3.7	2.8	1.8
5	1462	0.89	0.42	2.7	2.4	2.8	88.7	2.1
6	1462	0.81	0.32	1.4	6.0	9.9	80.9	0.6
7	1462	0.81	0.45	7.3	4.9	81.1	4.6	0.8
8	1462	0.82	0.35	4.2	81.5	2.4	9.9	0.6
9	1462	0.94	0.42	1.4	1.6	93.5	1.4	0.9
10	1462	0.76	0.22	76.5	1.6	2.3	17.6	0.6
11	1462	0.65	0.33	8.8	65.0	4.9	19.3	0.8
12	1462	0.94	0.34	93.8	1.5	2.1	0.6	0.7
13	1462	0.87	0.36	1.9	7.2	87.3	1.7	0.8
14	1462	0.44	0.22	44.1	21.2	21.2	11.8	0.6
15	1462	0.69	0.33	69.2	6.4	11.4	11.2	0.7
16	1462	0.55	0.26	5.2	55.1	28.8	9.0	0.8
17	1462	0.6	0.35	20.0	11.8	59.7	6.3	0.9
18	1462	0.66	0.45	14.9	66.3	6.9	9.8	1.0
19	1462	0.85	0.4	4.5	2.9	4.6	85.5	1.2
20	1462	0.46	0.2	37.1	6.6	45.8	8.1	1.3
21	1462	0.79	0.32	7.9	3.8	6.8	79.3	1.2
22	1462	0.59	0.43	58.7	7.1	14.6	17.0	1.7

Grade 3-5 (Form C-2) Reading Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1462	0.66	0.38	9.9	18.0	65.9	3.0	3.3
2	1462	0.91	0.34	3.6	90.6	1.9	0.8	3.1
3	1462	0.73	0.32	73.5	3.3	19.0	1.0	3.0
4	1462	0.92	0.44	2.0	1.0	2.1	91.8	3.2
5	1462	0.81	0.27	5.2	5.8	81.1	4.2	3.6
6	1462	0.66	0.46	65.7	6.2	6.6	17.7	3.6
7	1462	0.57	0.35	12.8	57.0	17.1	9.6	3.4
8	1462	0.8	0.44	7.2	80.0	7.6	1.1	3.8
9	1462	0.45	0.31	13.4	34.3	44.5	3.6	4.2
10	1462	0.75	0.42	5.1	9.4	6.2	75.0	3.9
11	1462	0.5	0.22	49.9	21.3	17.4	7.1	4.2
12	1462	0.49	0.44	21.6	9.6	48.8	15.5	4.4
13	1462	0.54	0.39	4.6	54.0	28.0	8.7	4.5
14	1462	0.8	0.45	79.6	5.3	3.4	7.0	4.5
15	1462	0.8	0.48	4.8	3.9	6.2	80.4	4.5
16	1462	0.66	0.46	12.4	6.0	10.3	65.7	5.4
17	1462	0.72	0.48	72.4	9.6	7.3	5.3	5.2
18	1462	0.56	0.32	8.9	10.6	56.0	18.4	6.1

Grade 3-5 (Form C-2) Writing Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1462	0.82	0.43	4.5	82.2	6.1	5.0	2.3
2	1462	0.92	0.44	2.8	1.7	91.7	1.0	2.6
3	1462	0.84	0.46	7.6	3.4	2.1	84.3	2.4
4	1462	0.81	0.36	2.2	4.5	81.1	9.5	2.7
5	1462	0.86	0.43	0.8	1.0	86.4	9.1	2.6
6	1462	0.8	0.47	7.8	5.5	3.2	80.2	2.7
7	1462	0.92	0.43	2.1	91.5	1.6	2.1	2.6
8	1462	0.81	0.5	8.8	81.3	3.8	3.4	2.7
9	1462	0.73	0.28	72.5	19.1	3.3	1.9	3.2

Grade 3-5 (Form C-2) Speaking Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	1462	0.8	0.29	20.1	79.9	-	-	-
2	1462	0.91	0.31	9.4	90.6	-	-	-
3	1462	0.94	0.29	6.0	94.0	-	-	-
4	1462	0.97	0.26	3.5	96.5	-	-	-
5	1462	0.98	0.28	1.9	98.2	-	-	-
6	1462	0.97	0.23	3.0	97.0	-	-	-
7	1462	0.95	0.24	5.3	94.7	-	-	-
8	1462	0.94	0.29	5.8	94.2	-	-	-
9	1462	0.9	0.33	10.3	89.7	-	-	-
10	1462	0.91	0.33	9.0	91.0	-	-	-
11	1462	0.85	0.42	3.6	22.8	73.7	-	-
12	1462	0.81	0.5	4.9	28.8	66.3	-	-
13	1462	0.68	0.44	9.5	9.2	16.8	30.0	34.6
14	1462	0.72	0.5	4.2	6.5	19.6	35.2	34.5

Grade 3-5 (Form C-2) Reading Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
19	1462	0.5	0.47	29.2	41.2	29.6	-	-

Grade 3-5 (Form C-2) Writing Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
10	1462	0.68	0.52	10.5	43.0	46.6	-	-
11	1462	0.51	0.58	8.5	20.5	38.4	23.7	8.9
12	1462	0.34	0.53	26.3	31.1	27.1	10.7	4.7

Grade 6-8 (Form D-2) Listening Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1448	0.91	0.29	91.2	2.7	2.4	2.0	0.2
2	1448	0.82	0.24	5.6	82.3	8.6	1.8	0.2
3	1448	0.9	0.18	89.6	3.9	4.3	0.4	0.3
4	1448	0.9	0.33	4.6	1.4	89.6	2.8	0.1
5	1448	0.72	0.32	10.5	72.1	8.4	7.5	0.1
6	1448	0.86	0.32	85.9	3.4	5.0	4.1	0.2
7	1448	0.78	0.31	5.3	9.9	78.1	5.0	0.1
8	1448	0.94	0.42	93.8	1.4	1.9	1.2	0.4
9	1448	0.91	0.38	1.5	90.6	2.6	3.6	0.2
10	1448	0.85	0.34	9.7	2.1	1.7	84.7	0.2
11	1448	0.73	0.34	2.7	8.1	73.0	14.6	0.1
12	1448	0.91	0.36	90.9	2.4	3.3	1.8	0.2
13	1448	0.91	0.35	4.3	90.8	2.0	1.2	0.3
14	1448	0.57	0.34	8.4	4.5	28.3	56.6	0.7
15	1448	0.72	0.32	2.8	72.2	15.3	7.5	0.6
16	1448	0.86	0.28	2.1	3.4	6.8	85.8	0.3
17	1448	0.84	0.26	8.6	3.3	2.5	83.6	0.4
18	1448	0.69	0.34	8.7	9.3	11.0	68.9	0.5
19	1448	0.76	0.38	76.4	6.2	13.5	1.5	0.8
20	1448	0.8	0.36	3.3	4.6	80.2	9.5	1.0
21	1448	0.65	0.16	10.6	9.6	65.4	12.0	0.8
22	1448	0.73	0.38	5.1	73.5	11.5	7.6	0.8

Grade 6-8 (Form D-2) Reading Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1448	0.74	0.26	2.6	5.8	74.5	14.2	3.0
2	1448	0.8	0.4	80.5	15.3	0.6	0.6	3.0
3	1448	0.91	0.49	90.6	1.9	2.1	2.2	2.9
4	1448	0.71	0.36	2.8	71.3	18.0	4.7	3.0
5	1448	0.86	0.47	1.7	4.3	5.5	85.6	3.0
6	1448	0.8	0.42	2.0	3.5	80.4	10.8	3.4
7	1448	0.69	0.38	69.3	12.6	10.5	4.5	3.0
8	1448	0.47	0.36	25.4	14.0	10.4	47.0	3.3
9	1448	0.53	0.34	23.1	8.4	12.2	53.0	3.1
10	1448	0.65	0.38	9.2	64.6	10.7	12.4	3.0
11	1448	0.7	0.45	14.1	9.3	70.2	2.9	3.3
12	1448	0.38	0.22	35.0	16.1	6.7	38.4	3.5
13	1448	0.76	0.43	9.8	5.9	76.2	4.4	3.4
14	1448	0.56	0.37	16.4	7.8	15.1	55.9	4.8
16	1448	0.55	0.38	10.5	19.1	55.3	9.7	5.4
17	1448	0.72	0.45	72.4	8.7	8.2	4.8	5.7
18	1448	0.58	0.4	8.4	58.4	10.3	16.9	5.9
19	1448	0.48	0.39	15.0	11.0	48.0	20.0	6.1

Grade 6-8 (Form D-2) Writing Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1448	0.92	0.45	91.6	0.9	2.9	1.0	3.7
2	1448	0.94	0.48	0.6	93.9	1.4	0.5	3.7
3	1448	0.83	0.48	7.3	83.2	4.1	1.2	4.1
4	1448	0.69	0.39	20.2	3.9	2.6	69.3	3.9
5	1448	0.73	0.35	11.4	8.5	2.9	73.4	3.8
6	1448	0.73	0.49	6.6	73.0	9.3	7.0	4.1
7	1448	0.87	0.54	2.4	2.6	3.7	86.9	4.2
8	1448	0.7	0.34	69.8	8.7	14.9	2.3	4.1
9	1448	0.42	0.25	9.1	7.3	41.9	37.1	4.5
10	1448	0.66	0.34	5.3	5.0	65.8	19.6	4.1

Grade 6-8 (Form D-2) Speaking Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	1448	0.9	0.4	10.2	89.8	-	-	-
2	1448	0.89	0.53	10.6	89.4	-	-	-
3	1448	0.92	0.52	7.7	92.3	-	-	-
4	1448	0.93	0.53	6.9	93.1	-	-	-
5	1448	0.92	0.49	7.7	92.3	-	-	-
6	1448	0.91	0.52	8.6	91.4	-	-	-
7	1448	0.89	0.42	10.8	89.2	-	-	-
8	1448	0.84	0.41	16.2	83.8	-	-	-
9	1448	0.8	0.38	20.2	79.8	-	-	-
10	1448	0.84	0.39	16.1	83.9	-	-	-
11	1448	0.79	0.57	9.7	23.6	66.8	-	-
12	1448	0.77	0.54	8.8	27.4	63.8	-	-
13	1448	0.65	0.54	10.8	10.5	18.7	28.6	31.5
14	1448	0.69	0.58	8.8	6.8	17.5	33.4	33.5

Grade 6-8 (Form D-2) Reading Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
15	1448	0.43	0.49	17.7	79.1	3.2	-	-
20	1448	0.36	0.53	26.3	19.8	40.7	9.7	3.6

Grade 6-8 (Form D-2) Writing Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
11	1448	0.64	0.59	14.9	41.8	43.4	-	-
12	1448	0.43	0.59	8.7	35.2	36.5	15.7	3.9
13	1448	0.37	0.55	20.7	32.6	29.0	14.1	3.7

Grade 9-12 (Form E-2) Listening Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1774	0.89	0.36	0.9	6.1	89.4	1.2	0.5
2	1774	0.89	0.34	6.4	88.7	1.0	1.3	0.6
3	1774	0.93	0.34	1.8	2.4	93.0	0.1	0.8
4	1774	0.83	0.43	4.2	6.9	82.9	3.5	0.3
5	1774	0.91	0.44	1.9	90.5	1.2	4.0	0.3
6	1774	0.9	0.4	1.6	4.3	1.1	90.4	0.4
7	1774	0.72	0.47	11.5	72.3	8.6	5.2	0.3
8	1774	0.88	0.44	3.6	87.8	2.8	3.2	0.4
9	1774	0.79	0.4	79.4	1.0	16.4	0.8	0.3
10	1774	0.89	0.4	5.2	2.3	89.2	0.9	0.3
11	1774	0.92	0.47	92.3	2.7	1.1	1.5	0.3
12	1774	0.75	0.24	1.2	75.4	19.8	1.2	0.3
13	1774	0.94	0.39	0.5	1.1	2.1	94.0	0.3
14	1774	0.8	0.37	4.1	12.4	80.3	1.0	0.2
15	1774	0.7	0.29	13.5	3.2	10.7	70.1	0.4
16	1774	0.91	0.4	1.2	2.0	3.2	91.0	0.5
17	1774	0.91	0.39	0.7	91.3	2.3	3.2	0.3
18	1774	0.75	0.37	74.7	9.5	6.0	6.9	0.7
19	1774	0.84	0.45	4.4	5.8	83.8	3.8	0.1
20	1774	0.52	0.37	9.2	22.0	13.7	52.3	0.6
21	1774	0.49	0.24	17.1	17.1	49.2	14.0	0.3
22	1774	0.6	0.35	12.7	60.3	11.8	12.8	0.3

Grade 9-12 (Form E-2) Reading Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1774	0.8	0.32	80.3	13.5	2.4	1.9	1.9
2	1774	0.91	0.42	2.2	3.3	1.5	90.9	2.1
3	1774	0.66	0.42	5.1	14.4	66.2	12.0	2.2
4	1774	0.51	0.33	6.9	28.0	11.7	51.2	2.2
5	1774	0.5	0.21	36.3	6.5	50.1	5.0	2.0
6	1774	0.85	0.45	4.0	7.3	84.7	1.9	2.0
7	1774	0.77	0.42	77.2	18.9	1.2	0.6	2.1
8	1774	0.76	0.38	1.1	18.5	76.5	1.9	2.0
9	1774	0.93	0.51	0.9	93.4	1.9	1.6	2.2
10	1774	0.57	0.28	57.2	13.8	11.3	15.6	2.1
11	1774	0.32	0.25	24.4	28.8	13.3	31.5	2.0
12	1774	0.67	0.21	2.9	6.8	67.3	20.8	2.3
13	1774	0.75	0.43	7.4	75.1	7.4	7.8	2.2
14	1774	0.86	0.45	2.7	3.3	5.5	86.2	2.3
15	1774	0.58	0.38	14.9	58.0	14.2	10.0	2.7
17	1774	0.81	0.51	3.7	3.5	8.3	81.3	3.3
18	1774	0.66	0.4	65.5	16.7	9.5	5.0	3.3
19	1774	0.62	0.4	62.0	12.4	16.1	5.6	3.8
20	1774	0.57	0.38	13.3	15.7	57.4	9.6	4.0

Grade 9-12 (Form E-2) Writing Items –MC

Item	N	p-value	PtBis	Percent Response Selected				
				A	B	C	D	Blank
1	1774	0.79	0.25	79.4	4.6	12.6	1.2	2.3
2	1774	0.89	0.37	2.8	3.0	89.3	2.5	2.5
3	1774	0.79	0.28	10.9	79.0	2.5	5.2	2.3
4	1774	0.88	0.37	87.6	6.2	1.4	2.4	2.4
5	1774	0.62	0.27	15.1	19.0	62.2	1.4	2.4
6	1774	0.75	0.43	74.7	12.2	7.1	3.5	2.4
7	1774	0.15	0.06	58.4	4.9	15.3	19.1	2.3
8	1774	0.84	0.46	3.0	84.3	7.3	2.8	2.4
9	1774	0.59	0.29	7.3	58.9	6.3	25.0	2.4
10	1774	0.69	0.34	9.2	2.7	16.6	69.2	2.3

Grade 9-12 (Form E-2) Speaking Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
1	1774	0.98	0.25	2.0	98.0	-	-	-
2	1774	0.96	0.26	4.4	95.6	-	-	-
3	1774	0.98	0.26	1.9	98.1	-	-	-
4	1774	0.98	0.27	1.8	98.3	-	-	-
5	1774	0.96	0.27	3.8	96.2	-	-	-
6	1774	0.93	0.32	7.0	93.0	-	-	-
7	1774	0.96	0.21	3.6	96.5	-	-	-
8	1774	0.9	0.23	10.1	89.9	-	-	-
9	1774	0.81	0.37	19.0	81.0	-	-	-
10	1774	0.78	0.22	21.5	78.5	-	-	-
11	1774	0.9	0.35	3.7	13.0	83.3	-	-
12	1774	0.86	0.41	4.5	19.3	76.2	-	-
13	1774	0.71	0.47	4.5	8.2	18.2	35.3	33.8
14	1774	0.69	0.44	6.7	9.4	18.3	33.9	31.8

Grade 9-12 (Form E-2) Reading Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
16	1774	0.64	0.5	21.3	29.0	49.8	-	-
21	1774	0.46	0.49	13.0	22.7	37.7	19.7	7.0

Grade 9-12 (Form E-2) Writing Items –CR

Item	N	p-value	PtBis	Score Point Distribution				
				0	1	2	3	4
11	1774	0.51	0.38	20.0	58.9	21.1	-	-
12	1774	0.46	0.56	9.1	25.7	41.5	19.3	4.3
13	1774	0.46	0.52	12.8	23.8	37.0	21.0	5.4

Appendix 2: MontCAS ELP Standards Setting Report

Setting Standards for the MontCAS English Language Proficiency (ELP) Assessment: Final Report

The MontCAS English Language Proficiency Assessment (ELP) is a modified version of an assessment developed for the Mountain West Consortium and designed to fulfill the requirements of ‘No Child Left Behind’ (NCLB) legislation. The MontCAS ELP assesses English proficiency in Listening, Speaking, Reading, and Writing, and reports scores in each of those language domains as well as in Comprehension (a combination of select items from the Listening and Reading test) and a total score, representing overall English proficiency. The MontCAS ELP was designed to assess the status of a student’s proficiency in English and to measure progress in attaining English proficiency.

As part of a contract with the Montana Office of Public Instruction (OPI), Questar assessment was charged with preparing standards setting materials as well as facilitating standards setting panels for the purpose of recommending cut scores that correspond to each level of English proficiency as defined by the Montana OPI. This document describes the standards setting process and resulting data. Prior to convening standards setting panels, an implementation plan was developed by Questar Assessment and presented to OPI for approval. Cut scores were recommended by panels using a procedure approved by OPI; panel recommendations for student performance by grade were presented to representatives of OPI for approval.

On February 28 through March 2, 2007, two panels, consisting of 26 Montana educators, were convened for the purpose of setting standards on the MontCAS English Language Proficiency (ELP) Assessment. The MontCAS ELP consists of forms administered in five grade spans: K, 1-2, 3-5, 6-8, and 9-12. One panel focused on the lower grades, K, 1-2, 3-5, and the second panel focused on middle and high school grades, 6-8, 9-12. Participants were chosen by OPI from a range of different stakeholder groups and were assigned to panels based on their experience with elementary or secondary education. Panelists clearly understood that their role was that of an *advisory group*- to recommend a set of standards. An agenda for the meeting is provided as Appendix A of this report.

The general methodology used for all sessions was an outgrowth of earlier “item mapping” procedures (Cizek & Bunch, 2007). This method, initially proposed by CTB/McGraw-Hill and termed the “Bookmark ProcedureTM” (Mitzel, Lewis, Patz, & Green, 2001; Lewis, Green, Mitzel, Baum, & Patz, 1998), was chosen for several reasons. First, it is currently the most widely used method for setting standards for high-stakes K-12 educational assessments and is used in the majority of statewide testing programs for which student performance standards are determined by panels. Second, it is well-suited for assessments, like the MontCAS ELP, that contain a mix of multiple choice and multi-point constructed response items.

Each panel member received an ‘ordered item booklet’, containing test items for the grade span under consideration. All items from both level 1 and level 2 forms within a grade cluster were included in the same booklet. A single test item was displayed on each page of the booklet and

pages ordered in terms of increasing item difficulty, as established in the concurrent Rasch item calibration based on the test administration in Idaho. Items were not separated by modality and constructed-responses items had a separate location in the book for each score point. Using the Bookmark procedure, panelists made “cuts” by placing markers in the books to indicate the item on which a student who could be characterized as minimally within one of the proficiency categories (e.g., just over the boundary of “proficient”) is more likely than not (i.e., with a probability greater than 0.50) to answer the item correctly. In many applications of this procedure, panelists are instructed to place the cut where the “minimally qualified” examinee has a 0.67 probability (more often expressed as 2/3 chance) of answering correctly. We tend to favor the 0.50 criterion for its simplicity for the judge. There is ongoing debate about which criterion should be used (see, e.g., Cizek & Bunch, 2007) but, at this point in time, both criteria are accepted.

Three rounds of cuts were planned for each grade span. In each round, panelists made cuts for each proficiency level by grade for each of the grades within the grade span under consideration. Following each of the first two rounds, panelists were shown frequency distributions and medians of recommended cuts and were given the opportunity to discuss the process. The second round was followed by impact data, i.e., the percent of students in each grade who would be placed in each proficiency level based on the median cuts assigned by the group. Although the items were ordered in terms of difficulties as established in the Idaho administration, the impact data were from the administration in Montana. The third round of cuts was accepted as the panelists’ final recommendations.

Standard setting results are shown on the following pages. Table 1 shows the median of the panelists’ final (round 3) recommendations, represented as scaled scores for each grade. The scaled score in each cell corresponds to the lowest score in the proficiency category represented by the column header for a test form and grade. Because each test form is unique, one can only compare scaled scores across grades where the same form is administered (e.g., grades 3-5).

Table 2 shows the percent of students, based on Fall 2006 test results, that would be placed in each proficiency category using the cut scores in Table 1.

Several adjustments to the panelists’ recommendations were proposed in order to create a more consistent pattern of results across the grades. A review of the distribution in Table 2 reveals fairly significant disparities over grades in the percent of students at different proficiency levels. This outcome is not uncommon when there are different panels working on different grade clusters. The adjustments were made to reduce these disparities. The shaded cells in Table 1 represent cut scores that were adjusted. Table 3 shows recommendations after adjustments and Table 4 shows the impact of those adjustments. The adjustments were considered by representatives of the Montana Office of Public Instruction (OPI) and adopted.

A summary of evaluation forms completed by the panelists is provided in Appendix B.

Table 1. Scaled Score Cuts by Grade and Form Based on Round 3

Grade	Form	Nearing Proficiency	Proficient	Advanced
K	A	363	396	425
1	B1&B2	345	374	421
2		373	408	466
3	C1&C2	361	384	408
4		374	397	425
5		387	403	464
6	D1&D2	367	392	411
7		367	392	416
8		370	392	444
9	E1&E2	370	395	424
10		373	396	424
11		376	400	435
12		376	400	435

Table 2. Percent of Students by Performance Level and Grade Based on Round 3 Cut Scores

Grade	Novice	Nearing Proficiency	Proficient	Advanced
K	23.2	38.9	30.6	7.3
1	9.2	21.4	56.2	13.2
2	8.5	11.2	64.4	15.9
3	4.4	13.7	44.8	37.1
4	4.0	15.2	49.1	31.7
5	6.8	12.9	76.2	4.1
6	3.5	27.5	46.1	22.9
7	5.5	19.7	54.8	20.0
8	7.8	21.0	68.8	2.4
9	3.1	31.9	62.4	2.6
10	2.3	26.6	65.9	5.2
11	4.8	31.9	61.4	1.9
12	4.4	32.1	60.9	2.6

Table 3. Recommended Adjustments to Scaled Score Cuts by Grade and Form

Grade	Form	Nearing Proficient	Proficient	Advanced
K	A	363	396	425
1	B1&B2	345	374	421
2		373	408	466
3	C1&C2	361	384	417
4		374	397	430
5		387	407	454
6	D1&D2	367	389	413
7		367	392	420
8		370	392	437
9	E1&E2	370	393	421
10		373	396	424
11		376	400	435
12		376	400	435

Table 4. Percent of Students by Performance Level and Grade Based on Recommended Adjustments to Round 3 Cut Scores

Grade	Novice	Nearing Proficiency	Proficient	Advanced
K	23.2	38.9	30.6	7.3
1	9.2	21.4	56.2	13.2
2	8.5	11.2	64.4	15.9
3	4.4	13.7	57.8	24.1
4	4.0	15.2	58.0	22.8
5	6.8	16.5	68.1	8.6
6	3.5	19.2	58.3	19.0
7	5.5	19.7	58.2	16.6
8	7.8	21.0	66.3	4.9
9	3.1	26.4	66.0	4.5
10	2.3	26.6	61.4	5.2
11	4.8	31.9	61.4	1.9
12	4.4	32.1	60.9	2.6

Table 5a shows, for each form and grade, the range of Total MontCAS ELP scaled scores corresponding to each proficiency level. Table 5b shows scaled score ranges corresponding to the proficient level in each of the language domains (Listening, Speaking, Reading, Writing) and Comprehension. Individual language domain tests do not include a sufficient number of items to reliably report more than two levels of proficiency. Procedures for establishing cuts in the language domains are detailed in Section 11 of the Technical Report.

Table 5a. Total MontCAS ELP Scaled Scores Corresponding to Proficiency Levels

Form	Grade	Total MontCAS ELP Levels			
		Novice	Nearing Proficiency	Proficient	Advanced
A	K	Below 363	363-395	396-424	425-
B1/B2	1	Below 345	345-373	374-420	421-
	2	Below 373	373-407	408-465	466-
C1/C2	3	Below 361	361-383	384-416	417-
	4	Below 374	374-396	397-429	430-
	5	Below 387	387-406	407-453	454-
D1/D2	6	Below 367	367-388	389-412	413-
	7	Below 367	367-391	392-419	420-
	8	Below 370	370-391	392-436	437-
E1/E2	9	Below 370	370-392	393-420	421-
	10	Below 373	373-395	396-423	424-
	11	Below 376	376-399	400-434	435-
	12	Below 376	376-399	400-434	435-

Table 5b. Language Domain MontCAS ELP Scale Scores Corresponding to Proficient Level

		Language Domain Proficiency Levels	
Form	Grade	Below Proficient	Proficient and Above
A	K	Below 98	98 and above
B1/B2	1	Below 91	91 and above
	2	Below 103	103 and above
C1/C2	3	Below 92	92 and above
	4	Below 99	99 and above
	5	Below 103	103 and above
D1/D2	6	Below 95	95 and above
	7	Below 96	96 and above
	8	Below 96	96 and above
E1/E2	9	Below 96	96 and above
	10	Below 98	98 and above
	11	Below 100	100 and above
	12	Below 100	100 and above

References

- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the meeting of the 1998 National Council on Measurement in Education, San Diego, CA.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Appendix A

Standards Setting Session Agenda

MontCAS English Language Proficiency (ELP) Assessment

Wednesday, February 28 – AM

8:00 – 8:30 **Continental Breakfast**

8:30 – 9:00 **Welcome, Introductions** (by Dept. staff member)
After Introductions, divide into separate rooms:
Gr. K-5 in one room; Gr. 6-12 in the other

9:15 – 10:30 **Orientation to Setting Standards**
Delimit the panel's activities – "Groundrules"
Agenda for the 3 days
What does it mean to set "performance standards"?
Overview of the general process of setting standards
Process of placing cut scores to segment a continuum of performance
Drawing a discrete cutoff (threshold students)
Errors of classification in any measurement process
Why multiple rounds are required
Keys to making good judgments

10:30 – 10:45 **Break**

10:45 – 12:15 **Definitions and Description of Performance Standards**

Performance Level Descriptors to be used at all grade levels – labels & Descriptors
*Note – the performance level descriptors describe what a student **SHOULD BE ABLE TO DO AT MASTERY** of that level, rather than at the entry to that level.
Making these general descriptors concrete for the specific grades
What does it mean for a student to be described this way – What can these students *do*? What do they *know*?

Wednesday – PM

12:15 – 1:00 **Lunch**

1:00 – 2:00 **"Experience" the Test**
Overview of framework for the MontCAS ELP
"Take" the actual assessment(s) on which standards will be set
Discuss the assessment – content, concerns, difficulty, assessed domains

2:00 – 2:45 **Orientation to the Specific Standard-Setting Methodology**
"Mechanics" of setting standards using "item mapping" procedure; judges' task
Features of the "item mapping" method – how it "works"
How materials are sequenced

2:45 – 3:15 Preparation for Round 1 of Judgments

Reminders of key issues
Distribute materials and orient panelists to use
What to do – how to indicate cuts
 Mechanics of filling in judgments
Rules for ratings – anonymity, independence, mechanics, security of materials, Day 2 overview

3:15 – 3:30 Break

3:30 – 5:30 (or until completion) First Round of Judges' Work

Panelists work independently, completing judgments for 3 grades (Judges turn in sheet/booklet for K before beginning Gr. 1-2; judges for Grades 6-8 complete all 3 grades concurrently). Judges turn in rating forms and leave for the day when completed

Thursday, March 1 – AM

8:00 – 8:30 Continental Breakfast

8:30 – 8:45 Review of Round 1 Issues and Problems

Questions/Observations of judges to the process in Round 1
Clarification of general issues and “mechanics” of the process

8:45 – 10:30 Feedback & Discussion of Round 1 Ratings

Feedback on Round 1 – Graphic portrayal of all panelists' ratings – by grade
Meaning of Round 1 ratings – distribution of cuts, median/mean cuts
Discussion of results across the three grade levels – do these make sense?
Discussion of WHY's for Round 1 (i.e., what led panelists to set their standards as they did? Problems, issues, confusions, rationales for preliminary standard)
Discussion of selected items or score points on extremes and near the middle of the Round 1 distribution of cuts
“Shaping” of panelists' considerations and judgments, focusing on critical considerations (threshold performance, “should vs. will,” PLDs, item mapping procedural confusions, construct issues)
Purpose of Rounds 2 & 3 – reflection, reconsideration, and comfort, not consensus
Student performance data by item by grade
What the data mean and why they are only minimally useful in setting standards
Reminder of key considerations

10:30 – 10:45 Break

10:45 – 12:00 (or completion) Round 2 of Judges' Work

Opportunity to reconsider and adjust Round 1 judgments

Thursday – PM

- 12:45 – 1:45 **Review of Round 2 Judgments**
Questions/Observations of judges on the process
Feedback and discussions much like that for Round 1
Anticipated statewide “impact data” by grade
Discussion of selected items or score points
- 1:45 – 2:00 **Preparation for Final Judgments**
Evaluation forms - returned directly to Dept. staff
Questions, reminders
- 2:15 – 3:30 (or until completion) **Final Round & Evaluation**
(panelists stay in area after completing work)
- 3:45 **Reconvene**
- 3:45 – 5:00 **Review and discuss *next* level of assessment & review PLDs**
- 5:00 – 5:15 **Prepare for First Round of Judgments for second assessment**
- 5:15 – 6:30 (or when finished) **First Round of Judgments – Second Assessment**
(Grades 3-5 for one panel; Grades 9-12 for the other)

Friday, March 2 – AM

8:00 – 8:30 Continental Breakfast

- 8:30 – 9:00 **Review of Day 2 Final Recommendations for the *First* Assessments**
Impact Data; consistency across grades
- 9:00 – 10:30 **Feedback & Discussion of Round 1 Ratings for 2nd assessment**
Feedback on Round 1 – Graphic portrayal of all panelists’ ratings by grade
Meaning of Round 1 ratings – distribution of cuts, median/mean cut
implications for statewide outcomes – (if to be presented)
Discussion of WHY’s for Round 1 (i.e., what led panelists to
set their standards as they did? Problems, issues, confusions,
rationales for preliminary standard)
Relationship between Round 1 cuts and final recommendations for 1st assessment
Discussion of selected items or score points on extremes and near the
middle of the Round 1 distribution of cuts
“Shaping” of panelists’ considerations and judgments, focusing on
critical considerations (threshold performance, “should vs. will,”
descriptors, item mapping procedural confusions, construct issues)
Student performance data by item by grade

What the data mean and how they are used in setting standards
Reminder of key considerations

10:30 – 10:45 **Break**

10:45 – 12:00 (or completion) **Round 2 of Judges' Work**
Opportunity to reconsider and adjust Round 1 ratings

Friday – PM

1:00 – 2:15 **Review of Round 2 Judgments**
Questions/Observations of judges on the process
Feedback and discussions much like that for Round 1
Anticipated statewide impact data – all grades
Discussion of selected items or score points
Convergence with/Differences from recommendations for other grades

2:15– 2:45 **Preparation for Final Judgments**
Evaluation forms
Questions, reminders
Wrap up/thanks – *Dept. staff*

2:45 – 4:00 (or until completion) **Final Round of Judgments & Evaluation**
(panelists depart as they finish work)

Appendix B

Summary of Panel Evaluation Forms—Collapsed across all 4 panels (based on complete sample of all panelists—48 completed Evaluation Forms)

Montana *English Language Proficiency Assessment* Standards-Setting Sessions Winter 2007 Evaluations Summary

1. Indicate the level of success of various components of the standards-setting session in which you participated.

<i>Component</i>	<i>Not Very Successful</i>	<i>Partially Successful</i>	<i>Successful</i>	<i>Very Successful</i>
Overview of the process of setting standards		4%	63%	33
Performance Level Descriptor review	2%	6%	60%	31%
Review of the actual <i>ELP</i> assessments		6%	65%	29%
Review of Round 1 results and interpretation		8%	52%	38%
Review of Round 2 results and interpretation			52%	48%
Group discussions of the panel	2%	15%	35%	48%
Data presentations before Rounds 2 & 3			46%	52%

Note: Percentages do not always total 100% because one or two panelists did not respond to every item.

2. Indicate the importance of each of these factors in making your cut-score

<i>Factor</i>	<i>Not Important</i>	<i>Somewhat Important</i>	<i>Important</i>	<i>Very Important</i>
Performance Level Descriptors		21%	42%	35%
Your perception of the assessment's difficulty	6%	10%	46%	38%
Your own professional experiences		10%	42%	44%
Your initial judgments (Round 1)	2%	35%	42%	21%
Group discussions of the panel	2%	6%	33%	58%
Item-by-item state data (prior to Round 2)		10%	58%	29%
Likely statewide impact data (prior to Round 3)		4%	52%	44%
Policy environment in the state	2%	23%	40%	31%
What students <i>would</i> vs. <i>should</i> be able to do	6%	4%	35%	52%

recommendations

Note: Percentages do not always total 100% because one or two panelists did not respond to every item.

3. I understood the task of recommending performance standards when I made judgments for:

	<i>Not Very Well</i>	<i>Moderately Well</i>	<i>Very Well</i>
Round 1	19%	46%	31%
Round 2		42%	58%
Round 3		15%	85%

Note: Percentages do not always total 100% because one or two panelists did not respond to every item.

4. I understood the *data* that were provided to the panel prior to:

	<i>Not Very Well</i>	<i>Moderately Well</i>	<i>Very Well</i>
Round 2	4%	27%	67%
Round 3		10%	90%

Note: Percentages do not always total 100% because one or two panelists did not respond to every item.

5. How confident are you with your *personal* classification of students at each proficiency level?

<i>Performance Level</i>	<i>Not Confident</i>	<i>Somewhat Confident</i>	<i>Confident</i>	<i>Very Confident</i>
Novice	2%	10%	38%	48%
Nearing Proficiency		11%	42%	45%
Proficient	1%	8%	48%	41%
Advanced	1%	10%	36%	50%

Note: Percentages do not always total 100% because one or two panelists did not respond to every item.

6. What strategies did you use to recommend *ELPA* performance levels?

- Facilitator knowledge. Documents (descriptors). List each item with my own descriptor—list N, NP, P or A for each round to help compare.
- State and personal definitions. Cultural bias. Where did students come from. Language used at home. Ruby Payne. Rewards. Subjects matter.
- Would vs. should, looking at all Montana kids, not just mine.
- Group discussion to see diversity of ELPA kids in Montana.
- Policy Environment and how results will be used.
- Ignored questions which seemed inappropriate or had poor “answer” choices.
- Began with advanced and novice first, then looked at proficient last.
- I really looked at what the state provided as the performance level indicators. It was difficult to find corresponding items on the test. I’m not sure that the test is assessing what is in the descriptors.
- I used all the information given to me at each round then I based on experience of my teaching years statistic classes; also listening to peer comments.
- Reading, writing, communication of student grade—vs.—community vs. State, actual test questions were they easy to understand or difficult to answer? Age (audience), type of question W, R, S, L, and the chart.
- —my own review of all the different kinds of kids I’ve seen.
- —review of percentages of kids who passed items.
- Experience. Montana English Language Proficiency Level Descriptions. Knowledge of LEP groups in Montana; distinction between foreign students vs. native people.
- At beginning, I selected at the high end of my range for the performance levels. At the end, I selected at the high end of the range discussed by the majority panel members.

- The strategies used went off the concrete standards we put on the wall, questions I asked teachers that taught at grade level, and those who administered the test information of their observations, etc. then compared this to Item Statistic Table.
- Background knowledge and experience with ELP students and prior experience with administering assessment.
- The “concrete” sheets developed.
- The skills needed per test items—looking for natural divisions.
- Review level descriptors; experience in giving the assessment; group discussions; data presentations.
- Difficulty of test question.
- The data after each round was very helpful. Discussion was invaluable.
- Picturing specific children that I tested; discussion; mean scores.
- Listening to the “experts,” those people with the most experience at K-2 level.
- Consideration of my experience teaching reading, my experience learning foreign languages, the experience of others in the room, and the experience of others who gave the test.
- All we talked about; personal experience.
- Common knowledge of geographical and cultural communities’ language vs. skill; majority vs. minority, native speakers and learners.
- I really listened to the other panel members and chose a couple of questions (in context to these surroundings) that really exemplified what a “proficient” student should do.
- Looked at data; considered group impact and listened to group discussions.
- Multiple readings; weighing opinions/experiences of others.
- I relied on: 1) My knowledge that level of test—very easy according to “P” values; 2) Knowledge of population of students; 3) Knowledge of language acquisition; 4) Understanding of what LEP means.
- Listening to others’ experience; going back and rethinking then make decision.
- Review of p values; review of impact data; watch how the facilitator commented on process.
- Past experience; wide teaching experience in state.
- Performance descriptors and what skills the questions were really asking for.
- Looked at the questions, levels, and advice of other teachers.
- Group discussion; correct statewide.
- Use of concrete ideas—posted; discussions; looking at the skill each item needed to be answered correctly.
- Strategies/resources used: MT performance levels; MT EZP level description; wall charts; group discussion.
- Co-panelist’s discussions; personal experience; data review.
- Personal experiences with the test administration; group discussions, particularly from teachers at that level; place nearing proficiency and advanced first.
- Understanding the “should be able to” of a proficient child at each grade level, plus prior knowledge and experience
- Group input; performance descriptors.
- My background, other “experts” in the room discussing these and looking at what Montana standards, etc. to help us picture what each student should look like at that level.

- Personal experience and understanding of grade levels performance expectations.
- Where I felt we needed to raise our level; I voted high on rounds 1 & 2; then on 3rd round looked at where the median was and voted just a little higher.

7. What effect—positive or negative—did your experiences during the first two days of these sessions have on the judgments you made on the *final* set of recommendations?

- Positive—gained better understanding of testing questions, performance levels, other programs.
- Negative—sitting so long.
- Positive—meeting teachers from around the state.
- Negative—Cultural bias—left out—how to correct.
- Pos: working with a very knowledgeable group of people. Opinions expressed during discussions.
- Pos: facilitator
- Pos: Interesting and informative discussion and debate over standards and test items.
- Neg: Some of that debate seemed to take on a hostile and angry tone—I was uncomfortable during these times and I don't think that this was productive. It was actually counter-productive.
- Positive—I realized that my perception of the test initially was accurate—the test was easy for most students.
- Negative—frustration with questions chosen that I felt weren't well-worded but having to base standards upon them.
- Pos: I think that the group will come to a good recommendation.
- Pos: I appreciated the extent that the facilitators helped us with the process. The information provided and explanations were very positive.
- Pos: It was very interesting and educational to be in the session with wonderful educators with same/familiar teaching environments.
- Neg: The only negative was the distractions from others "talking" when presenter was informing
- Pos: The moderator was very good—funny—still shaping our participation—using data—cuz I use the number to help me recognize reality of performance.
- Positive—Pleasant facilitator. Interesting process
- Negative—Realizing how much time was wasted writing this test, piloting it, giving it, because it turned out to be too easy—a badly done assessment.
- Very, very stressful time of year to take a 3-day "time out" to help set standards.
- My students' test scores will suffer.
- Positive—the input from the other teachers I found quite valuable. Their experiences giving the tests, what knowledge working with students at that particular grade level. We were kept on track or put back on track very professionally.
- Pos: the group discussions, and input from others, and facilitator keeping the group on task were all positive.
- Neg: covering six grades in 3 days (time frame) was negative aspect.
- Neg: The inability for some people to understand that this test needed to have cut scores made for all Montana students not just the students they teach.

- Positive—the most positive was the item by item discussion. I learned much from others who gave the assessment
- Negative—it was a difficult environment in which to discuss item and experience.
- Positive—great experience in learning how to go about setting test standards.
- Negative—found the test to be not very helpful in setting standards
- Positive—discussion and data.
- Negative—time spent discussing minute populations.
- Neg: not having data for 1st and 2nd graders split out; having the test de-qualify 2nd graders who are not truly proficient, but showed as being so on the test.
- Pos: Group discussions, learning how to set standards.
- Neg: Some people not listening to the rest of the group's reasoning.
- Positive—understanding of what test and results are like
- Negative—do not have a lot of confidence in the actual test itself.
- Positive—hearing the discussion amongst educators, sharing ideas.
- Negative—thinking about the impact on student self-esteem of a low score, considering how limited any test is in its ability to measure what it is intended to measure.
- Pos: shared discussions with others.
- Neg: moved too slow—sometimes bogged down with personal issues.
- Did the best, I thought, contributed to discussion and decisions.
- I believe I learned more and became more confident as the process progressed!
- Definitely easier as process continued.
- Allowed me the time to be more reflective. Also I understood the terminology, policies, etc. so much better and was able to provide more solid and reflective responses.
- Refining process; coming together; coming to agreement.
- Took peer input and made minor adjustments, positive feedback, and experienced people.
- Negative—test was not very well-created to assess our students.
- Positive—great discussions.
- Positive in all ways.
- I felt I learned much during the first two days. The final set of recommendations was easier to do because I learned the process.
- Made it easier and clearer; more confident.
- My realization last night that—50% of the kids missing an item happened at item 90 of 104 which left us placing 6-9 cuts in those 14 items.
- Positive—All recommendations and rationales were great.
- Negative—one person going back to a task that had been completed and discussed earlier.
- More confidence on how I set classification levels.
- As the sessions progressed, I learned more from the “experts” on the panel, the discussion was enlightening and essential, and I hope you take in consideration the hard work put into this.
- Performance level descriptors.
- The experiences of days 1 and 2 had everything to do with my choices on day 3; I learned a tremendous amount about arguing my point with others who were just as passionate and dedicated as I am

8. What were the most positive *and* negative aspects of your participation in these standards-setting activities?

- I had never participated in a process like this before. I know I'll be looking at my students differently in terms of what they know/don't know or can/can't do.
- I thank you for allowing me to be part of the process. More people (not just educators) should have a better understanding of the whole process.
- Thanks, enjoyed very much. Especially your humor, that's my learning style and I always feel comfortable when everyone else can share a laugh—"life is too short and one without laughter is very sad!" Take care and God bless.
- The locale was an embarrassment. The work rooms were not comfortable. The tables were ugly—not even covered. The snacks were unhealthy. The overall ambiance for a 3-day workshop was unpleasant and Spartan. I felt that this hotel was not an appropriate choice for the sessions, and it gave a bad impression to the TASA people. What must they think of Montana?!
- Very educating and very enlightening. The most valuable process was the setting of descriptors of what is expected of at each level. This information would be very valuable for all teachers across the state to prepare for the test and assist their students to become proficient learners.
- I felt the facilitator at some points tried to sway the judgments. Example: "Should a 3 on an open-ended question be in Advanced? It seems that it should be in Proficient."
- Good experience, but process was very frustrating because of the test. The test was not very well written in a way that I felt could help us do a better job of setting the standards.
- I felt that 2nd grade was short-changed by the limitations of the test. I do not believe that only 17% of our second graders are not proficient English-language learners.
- Crowded meeting room; not a great hotel; last minute organization; teachers like to talk.
- Scared to be the deciding factor that will determine school funding vs. necessary programs that assist and promote important programs.
- Positive—facilitator; members of the panel
- Negative—facilities
- Positive—I realized that test was too easy and we needed to up the standards.
- Positive—more informed to make a better decision.
- The most positive aspect was that we had plenty of time to make decisions.
- The most negative aspect was that the jumps in P values made it very difficult to make out points accurately.
- Better understanding of student expectations.
- There was some confusion caused by the differences between LEP—Native Americans and LEP foreign students. I think our state's LEP population is not clearcut. N.A. educators didn't seem to be evaluating this test as a language test.
- Getting together in this size setting and working as we did.
- Positive—Discussions today were more like discussions rather than angry arguments like yesterday.
- Positive—group discussions; chances to comment about the testing inconsistencies.
- Equal.

- Negative—did not think the actual test is doing its job for Montana students; not much confidence in the test.
- Positive—learned a lot about the test and process.
- Negative—test was not very well-created to assess our students.
- Positive—great discussions.
- Positive—group discussion.
- Negative—bogged down in a poor test.
- Pos: Learned a lot!
- Neg: Disappointment in the test.
- Positive—moderator: Shirley; the group with whom I worked.
- Negative—the hotel, meeting rooms.
- Positive—sharing of information; reinforcement of the expertise in this state.
- Negative—people who didn't listen.
- I had hoped to address the issues of how low the test was for Montana's English speaking Native kids—that didn't happen.
- Pos: group dynamics, getting to know people.
- Neg: members of group off task
- As the sessions progressed, I learned more from the "experts" on the panel, the discussion was enlightening and essential, and I hope you take in consideration the hard work put into this.
- Negative—thinking about the futility of giving such a broad spectrum test (grades 3-5, developed for a broad audience) with the possible intent of basing funding decisions on it.
- Positive—Most was positive. Broadened my understanding of statistics, testing, teaching, my state, etc. etc. etc.
- Positive learning experience.

9. Use the space below to make any additional comments about the process or your experience. *Thank you for taking the time to evaluate the sessions.*

- I really enjoyed working with Mike.
- Thanks, Mike! You were great!
- I love the process—brainstorming informative.
- This was an incredible experience that has provided me with additional insight into my own students and their performances and abilities.
- Facilitator was very good 9-12.
- Safe trips home.
- Need better accommodations.
- Glad I came. Learned a great deal and hope to continue involvement.
- Thank you for the opportunity to be trained.
- Thank you, Sheila!
- Excellent opportunity to meet and learn from the remarkable educators that work in Montana.

Adapted from Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In Cizek, G. (Ed.) Setting performance standards: Concepts, methods and perspectives, Mahwah, NJ: Lawrence Earlbaum Associates.

Appendix 3

Mountain West Assessment Consortium Foundation Document

Introduction

The *Mountain West Assessment Consortium Foundation Document* is part of a response to the No Child Left Behind Act (NCLB) of 2001 that mandates assessment of English language learners' progress in attaining proficiency of academic English. Since regular state assessments may not accurately reflect the gains English language learners have made in attaining English proficiency, the Mountain West Assessment Consortium has developed an English language proficiency assessment to serve a dual purpose: to measure students' language proficiency and to measure students' progress toward meeting state standards. Through the development and administration of this assessment, Mountain West Consortium states will satisfy the NCLB requirements for monitoring the development of English proficiency of the English language learners in their public schools.

The *Mountain West Assessment Consortium Foundation Document* describes the elements of language proficiency that are the basis for the Mountain West Assessment Consortium's English Language Proficiency Assessment. The purpose of the assessment is to gauge English language learners' progress in learning to listen to, speak, read, and write in the English language. The assessment follows a developmental progression across and within distinct grade spans. It is based on five communication standards recognized as the linguistic underpinnings of language: phonology, morphology, vocabulary, syntax, and function. The standards have been further detailed in benchmark performance descriptors.

Standards and benchmark descriptors are common elements of any framework that describes what students should know and be able to do. Standards are like umbrellas; they are broad-based, encompassing a set of related skills and/or knowledge bases. Benchmarks are more specific statements that describe discrete tasks students will perform in order to demonstrate knowledge or skills within a standard. For example, under the vocabulary standard in reading, one benchmark descriptor is, "*Reads and understands common idioms.*"

The Mountain West Assessment Consortium English Language Proficiency Assessment includes separate modules for children at these grade spans: kindergarten through early first grade; mid-first grade through second grade; third grade through fifth grade; sixth grade through eighth grade; ninth grade through twelfth grade. Within each of these designated grade spans, assessment items have been developed to evaluate growth in English language acquisition across three broad developmental levels: early acquisition, intermediate, and transitional. The assessment battery modules include test items at each of the three developmental levels across the four modalities of listening, speaking, reading, and writing.

It is important to emphasize the breadth of these developmental levels and to recognize that they are not proficiency levels. The developmental levels of the standards are intentionally broad; they are used simply to make general classifications of test items within the assessment. Proficiency or performance levels specify what a student has achieved or demonstrated *relative to a set of standards*. There may, in fact, be as many as five distinct proficiency levels within these three broad developmental levels. Proficiency or performance levels are determined through standard-setting activities that yield cut-scores within the total range of test scores. There are several ways to determine proficiency levels, and each state that elects to use the Mountain West Assessment Consortium English Language Proficiency Assessment will apply its own process to determine proficiency levels.

Benchmarks have been grouped within five standards to reflect the dimensions of communicative competency:

- Phonology/Orthography standards are used to evaluate students' progress in understanding and correctly manipulating the sound system of English.
- Morphology standards are used to evaluate students' progress in understanding and using the rules of English word formation.
- Vocabulary standards are used to evaluate students' understanding and appropriate use of English words and phrases (semantic knowledge).
- Syntax standards are used to evaluate students' progress in understanding and using the rules of English sentence formation.
- Function/Discourse standards are used to evaluate students' ability to use and comprehend English in various oral and written contexts.

Since elements of some standards must be in place before others develop, the application of these five language standards varies across both grade spans and developmental levels. For example, phonology benchmarks are generally addressed more extensively at the early acquisition level than at intermediate or transitional levels. In addition, the requirements for competency in the four modalities (listening, speaking, reading, and writing) vary so that one modality may emphasize some standards over others. For example, expectations for syntax use are more pronounced in the language production modalities of speaking and writing. Similarly, assessment of function/discourse skills is addressed in greatest depth at the transitional level.

All of the standards and benchmarks included in this document are addressed in the assessment. The majority of the benchmarks are addressed in specific assessment tasks. Other benchmarks are addressed indirectly through holistic acts of listening, speaking, reading, or writing. In the receptive processes of listening or reading, acquisition of some benchmarks is inherent in demonstrations of comprehension of the language presented. Holistic scoring rubrics have been developed to encompass such benchmarks in the language production modalities of speaking and writing.

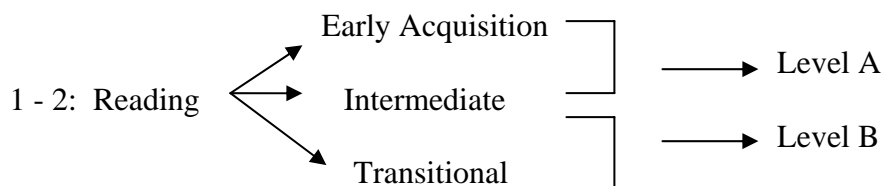
The order in which progress across the four language modalities is assessed also reflects a developmental perspective. The modalities generally considered informal - listening and speaking - precede assessment of the more formal language modalities of reading and writing. Moreover, since a degree of language comprehension generally precedes language production, receptive language skills are addressed before production skills in both informal

and formal order in the assessment. Thus, listening skills are assessed first, followed by speaking, reading, and writing skills in that order.

The developmental continuum is also reflected in this assessment in the degree to which language is decontextualized. At the early acquisition level, care has been taken to provide directions that are simple and concrete. Demonstration and practice items are also provided to help students understand what is expected of them. In addition, language in the test directions for intermediate and transitional level items begins to approximate the language found in mainstream assessments.

Spoken English proficiency is assessed in one-to-one settings and all K-1 assessment modules are administered individually. In response to pilot test feedback, all other modules of the Mountain West Assessment Consortium English Language Proficiency Assessment have been designed for group administration. However, the assessment of early acquisition level benchmarks and some intermediate level benchmarks is administrator-led (i.e., the test administrator reads directions and questions to the students). Assessment of intermediate and transitional level benchmarks (Level B; see below) is conducted in typical large-scale assessment format.

The Listening and Speaking assessment modules are designed to be administered in their entirety; each module begins with tasks reflecting early acquisition benchmarks and proceeds through tasks reflecting intermediate and transitional level benchmarks. With the exception of the K-1 measures, Reading and Writing assessment modules are designed so students take either Level A or Level B, but not both. Level A includes assessment tasks that progress from early acquisition benchmarks through early intermediate level. Level B encompasses assessment of intermediate and transitional level benchmarks. An illustration of these relationships appears below. A locator checklist is provided to assist test examiners in determining which test level is appropriate for each student.



Appendix 4

Secure Materials Check-In

The accounting of secure materials is done via a secure numbering system. Each secure material (test booklets, prompt books, Listening test CDs, and examiner manuals) will have a unique **secure ID number** assigned to it during the production process. This number and the associated document are then assigned and tracked to a receiving district and/or school. After testing, the district and/or school returns the document and Questar confirms receipt through their Secure Material Check-in process. To make this process as automated as possible, the ID numbers appear on the materials in a human-readable form and a barcode format. Each and every secure material barcode is scanned (or hand-entered if not scannable) and checked against the original shipping distribution file in our TestPath system to insure all materials have been returned. Upon receipt and scanning of all secure materials from districts and school, we complete a *Missing Secure Materials Report*.

For return of materials, Montana System Test Coordinators are instructed (instructions are in the Training Presentation, in the Test Coordinator Guide, and on the System ID Sheet) to place examiner manuals, listening CDs, unused answer documents, and unused test booklets in the bottom of their boxes and topped these with the goldenrod-colored divider sheet. Then Test Coordinators are instructed to pack the used non-scannable test booklets. Next the Test Coordinator packs in each completed scoring services envelope (containing completed answer documents and examiner identification sheet), organized by school (school identification sheet and voided barcode forms top each schools scoring services envelopes). When Questar opens the boxes, the first school's school id sheet, voided barcode label forms and scoring services envelopes are on top. Below is an overview of the steps.

Secure Materials Check-in Process. The check-in of secure materials is a critical step in maintaining the security and integrity of state testing programs. The Check-in process consists of four steps; material preparation, scanning, validation and storage.

Secure Material Preparation

Once materials are received at Questar, the materials are separated. Operators check-in the secure materials for answer documents and these are routed to answer document processing. They also organize the secure booklets for the scanning step. The materials are placed on carts as they come out of the boxes. As secure materials are stacked on the carts they will be checked for answer documents. If answer documents are found mixed in with the secure materials or actually slipped into a secure material booklet the answer documents are brought to the attention of the Operations supervisor.

Secure Material Scanning

Two check-in operators work together during this step. Each barcode is scanned twice (the first scan is for initial entry into the TestPath system and the second scan serves as a quality control check) and the documents are placed into storage boxes. The storage boxes are labeled with a specific box number that will be used in TestPath as a tracking number for the grouping of secure materials assigned to that box. During scanning, the document barcodes are stored in TestPath to be compared later to the original secure materials distribution data file. [On the system and school packing lists the ranges (or a single barcode if the case) of barcodes are provided. On the system identification sheet, the Test Coordinator is reminded

that materials have security serial numbers and to reference school and system packing lists to account for these materials as they pack items up.]

Secure Material Validation

Next is the validation step using TestPath. The validation step consists of processing each storage box through a series of checks in TestPath. It is during this step that the secure material ID numbers read via the barcodes in the scanning step are compared to the original distribution file. The checks involved in the validation process compare the ID numbers between the two scans conducted during the scanning step, ensures the scanner read a correct barcode format and compares the ID number to the distribution data file to make sure that number was actually assigned to a district and/or school. If these checks are all correct the box is validated “clean”. If the checks are not clean, the box is rechecked and a Secure Materials Validation Error Report is produced by TestPath. After this step, boxes are prepared for storage.

Secure Material Storage

The storage boxes are taped closed and stacked on pallets with the box labels facing out and taken to storage.

MontCAS

(Montana Comprehensive Assessment System)

English Language
Proficiency Assessment

2006-2007

Score Reports Interpretation Guide



Linda McCulloch, Superintendent

Montana Office of Public Instruction

PO Box 202501

Helena, Montana 59620-2501

www.opi.mt.gov

Contents

- 4 Overview**
- 6 Understanding the Individual Student Report**
- 8 Understanding the Parent Report**
- 9 Understanding the School Roster Report**
- 10 Understanding the Summary Report**
- 11 Using MontCAS ELP Results**

Overview

The purpose of this guide is to assist educators and other stakeholders with understanding, interpreting, and using the results of the Montana English Language Proficiency Assessment. The MontCAS ELP is administered statewide to all Limited English Proficient (LEP) students.

The guide includes information on

- how and why the MontCAS ELP was developed,
- how the assessments are designed,
- how student performance is scored,
- how performance standards were determined,
- how assessment results are reported, and
- how results can be used to improve programs, instruction, and student performance.

Purpose of the MontCAS ELP. The annual assessment of LEP students in Montana fulfills a requirement of the No Child Left Behind Act of 2001. One objective is to measure individual student’s progress in achieving proficiency in speaking, listening to, comprehending, reading, and writing English. A second objective is to measure in districts participating in Title III the success of language development programs in achieving adequate student growth in English proficiency.

Development of the MontCAS ELP. The MontCAS ELP is an edited version of the English Language Proficiency test developed for the Mountain West Consortium, of which Montana was a member. The first administration of the MontCAS ELP occurred in the fall of 2006. Using the data from this administration, psychometric work was completed by Questar Assessment, Inc. for the purpose of creating a score scale for each of the domains and for the total test. In February 2007, a panel of Montana educators met to set standards for the MontCAS ELP in the

form of cut scores for each proficiency level by grade. The 2007 MontCAS ELP score reports are the result of this process.

Structure of the MontCAS ELP. The MontCAS ELP is comprised of tests in four domains—Listening, Speaking, Reading, and Writing. Scores are reported for each of these domains, as well as for Comprehension. The Comprehension score is calculated using a subset of Listening and Reading items.

The MontCAS ELP is administered by grade span.

Grade Span	Form
K	A
1-2	B1 or B2
3-5	C1 or C2
6-8	D1 or D2
9-12	E1 or E2

In all grade spans, except for K, there are two separate Reading/Writing test forms, a Level 1 form intended for Beginning students and a Level 2 form intended for more proficient students. Having separate forms centered on two different ability levels made it possible to shorten the Reading and Writing tests. The Speaking and Listening tests, on the other hand, are the same for all students within a grade span. Note that no “mixed” scores can be reported: if, for example, a student took both B1 and B2 test forms, results have been reported for only one form.

Reported Scores. Student performance in each of the five language domains is reported in terms of raw score, scaled score, and proficiency level. Student performance on the overall (Total MontCAS ELP) test is reported in terms of raw score, scaled score, and proficiency level.

Raw Scores. The raw score is the total number of correct answers on multiple-choice items plus the number of points earned on open-ended items. Raw scores on the MontCAS ELP can only be compared for the same domain and the same test form. For example, a Form B1 raw score cannot be compared to a Form B2 raw score.

Note: The Writing raw score for (Kindergarten level) Form A was calculated as follows: 1 point was allocated for each skill on the Writing Checklist that the student "does most of the time" or of which they "demonstrate mastery." Thus, the Writing Checklist generated a maximum raw score of 22 points.

Scaled Scores. Scaled scores are derived from raw scores and provide results for alternate forms (e.g., B1 and B2) on a common scale. MontCAS ELP scaled scores can be compared for the same domain and the same grade-span test (A, B, C, D or E). For example, all Form C Reading scaled scores can be compared, regardless of whether the student took the C1 or the C2 Reading test. However, Form C scaled scores cannot be compared to Form D scaled scores.

Total MontCAS ELP Proficiency Levels. For the total score, four proficiency levels are reported: Novice (N), Nearing Proficiency (NP), Proficient (P), and Advanced (A). These are based on the total scaled score and provide a holistic estimate of the student's English proficiency. It is important to note that students at the same overall Proficiency Level may have different profiles of competence across the language domains.

Domain Proficiency Levels. Within each domain, two proficiency levels are reported, based on the student's scaled score: Below Proficient (BP) and Proficient or Above (PA). (Individual language domain tests are not long enough to reliably provide more than two levels of proficiency.)

Incomplete Testing. Students were required to take all four language domain tests. If a student did not take one or more of the domain tests, the reports will show dashes in place of scores for that domain. The reported Total MontCAS ELP score is based on the domain tests for which there are scores. Thus, if a student failed to take the Speaking Test for whatever reason, the Total MontCAS ELP score will be based on a raw score of zero in Speaking. The reported Comprehension scores—which are based on a subset of Listening and Reading scores—will be affected in the same way if the student failed to take either the Listening or Reading Test.

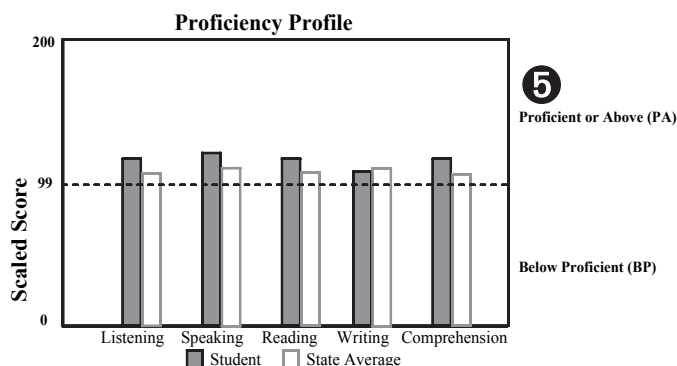
Individual Student Report

Student	HELMICK, FELICIA J
School	ABC School
System	ABC System
Grade	Grade 4
Test Form	C2
State Student ID:	987600032
Birth Date	07/19/1997
Gender	F
Test Date:	Fall 2006

The NCLB Act of 2001 requires an annual assessment of English language proficiency for students identified as limited English proficient (LEP). The purpose of the assessment is to measure students' progress in achieving proficiency in academic English. The MontCAS English Language Proficiency (ELP) Assessment measures proficiency in listening, speaking, reading, writing, and comprehension (domains). The comprehension score is a composite score based on the listening and reading sections. **Novice** students are beginning to participate in oral and written interactions of learned information to socialize, produce, and obtain information. **Nearing Proficient** students demonstrate partial mastery of oral and written interactions of learned information to socialize, produce, and obtain information. **Proficient** students demonstrate competent skills in oral and written interactions of learned information to socialize, produce, and obtain information in order to participate in academic work. **Advanced** students demonstrate exceptional skills in oral and written interactions of learned information to socialize, produce, and obtain information in order to participate in academic work.

Total MontCAS ELP (Max RS=83)	Raw Score	Scaled Score	Proficiency Level
	71	431	Advanced (A)
State Average Scaled Score		413.4	

Score Summary			
Test	Raw Score	Scaled Score	Proficiency Level
L Listening (Max RS=22)	20	117	PA
S Speaking (Max RS=22)	21	121	PA
R Reading (Max RS=20)	17	117	PA
W Writing (Max RS=19)	13	108	PA
C Comprehension (Max RS=39)	34	117	PA



Legend: RS: Raw Score; Max RS: Maximum Possible Raw Score; SS: Scaled Score; -- indicates test not taken

BP = Below Proficient **PA** = Proficient or Above

1 Test Form. Test forms are identified by a letter-number combination. The letter (A, B, C, D, or E) specifies the grade-span; the number specifies the difficulty level of the form (1 is for LEP students with beginner or novice skills in English; 2 is for the more proficient students). Note that the Speaking and Listening sections are identical; only the Reading and Writing sections are different on the Beginner (1) and Intermediate/Advanced (2) versions of the form. The exception is grade K (Form A), which does not have separate ability-level forms.

2 State Student ID. The state student ID is a unique number that is assigned to every student who receives educational services from a public school in Montana. This number follows the student from school to school throughout his or her K-12 career. The ID consists of 9 randomly generated digits, with no leading zeros.

3 The Raw Score is the total number of correct answers on multiple-choice items plus the number of points earned on open-ended items. A raw score can only be interpreted within the context of a given test form. Raw scores cannot be used to compare performance on different test forms. Scaled scores or scores derived from scaled scores should be used for those comparisons.

4 Scaled Scores are derived from raw scores and provide results for alternate forms (e.g., Forms B1 and B2) on a common scale. Scaled scores can be used to make comparisons among students and over time. However, scaled scores cannot be compared across test levels (e.g., B vs. C), or across different tests (e.g., Listening vs. Reading). To compare across different test levels, scaled scores must be converted to Proficiency Levels.

5 The Proficiency Profile summarizes ability across the language domains. The solid bars show individual ability, the striped bars show average ability statewide. The height of the solid bars shows how ability differs by language domain. The dotted line in the middle of the Proficiency Profile chart marks the cut score between the Below Proficient (BP) and the Proficient or Above (PA) levels, allowing you to see where student ability falls with respect to this criterion. Finally, comparing the height of the solid to the striped bar allows you to see how the test performance for this student measures up to performance statewide.

6 Proficiency Levels provide a holistic estimate of the student's English proficiency.

In general terms, the levels are:

Novice (N) – Students are beginning to participate in oral and written interactions of learned information to socialize, produce, and obtain information.

Nearing Proficiency (NP) – Students demonstrate partial mastery of oral and written interactions of learned information to socialize, produce, and obtain information.

Proficient (P) – Students demonstrate competent skills in oral and written interactions of learned information to socialize, produce, and obtain information in order to participate in academic work.

Advanced (A) – Students demonstrate exceptional skills in oral and written interactions of learned information to socialize, produce, and obtain information in order to participate in academic work.

The results of your student's English Language Proficiency Assessment are shown in this report by raw score, scaled score and performance level.

A **Raw score** refers to the number of points a student has earned for a particular test. Raw scores should not be compared across language domains. A maximum raw score is shown for each language domain and the Total MontCAS.

Scaled scores are derived from raw scores and permit comparisons between level 1 and 2 forms (e.g., Form C1 and C2) within a grade cluster. Scaled scores range from 0 to 200.

Performance levels describe a student's performance on the MontCAS ELP assessment and are based on the total scaled score. The MontCAS ELP reports four performance levels for the total score (N, NP, P, A), which are organized into two groups for each domain (BP, PA). These performance levels are described in more detail on the back cover.

YOUR STUDENT'S RESULTS

The following charts reflect your student's raw score, scaled score, and performance levels on the English Language Proficiency Assessment.

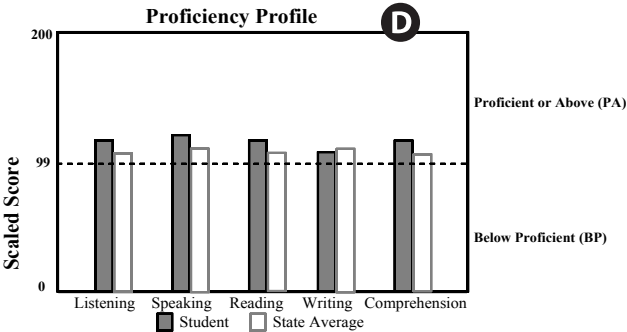
Total MontCAS ELP. This table indicates your student's overall performance on the assessment. In addition to information on your student's performance, state results are included for comparison. The score summary and proficiency profile on the next page illustrate more detailed information about how your child performed in each domain.

Total MontCAS ELP (Max RS=83)	Raw Score	Scaled Score	Proficiency Level
	71	431	Advanced (A)
State Average Scaled Score 413.4			

Score Summary. The Score Summary chart provides your student's results for each of five components of the ELP assessment: Listening, Speaking, Reading, Writing and Comprehension. The maximum raw score (Max RS) is indicated for each component. For example, the maximum raw score (Max RS) that could be earned for the Listening test was 22 points.

Score Summary			
Test	Raw Score	Scaled Score	Proficiency Level
L Listening (Max RS=22)	20	117	PA
S Speaking (Max RS=22)	21	121	PA
R Reading (Max RS=20)	17	117	PA
W Writing (Max RS=19)	13	108	PA
C Comprehension (Max RS=39)	34	117	PA

Proficiency Profile. The profile indicates your student's performance in relation to the proficiency levels and to the State Average.



Legend: RS: Raw Score; Max RS: Maximum Possible Raw Score; SS: Scaled Score; -- indicates test not taken
BP = Below Proficient PA = Proficient or Above

A customized parent report was generated for each LEP student who participated in the fall 2006 MontCAS English Language Proficiency (ELP) Assessment. This report was based on the school level individual student report and should be shared by classroom teachers during parent-teacher conferences or other interactions with parents. The report includes detailed results of a student's ELP test performance, including raw scores, scaled scores and performance levels, in each language domain and for the total MontCAS ELP. The proficiency profile permits a comparison of student ability across the language domains and in comparison to average performance across the state.

Section A provides an explanation of terms – raw score, scaled scores, and performance levels – used in the Parent Report.

Section B shows the student's overall performance on the assessment in the Total MontCAS ELP table. The

student's total raw score, scaled score, and proficiency level are provided, along with the Average State Scaled Score for comparison.

Section C provides more detailed information about student performance in the Score Summary chart. The chart shows student results for each component of the ELP assessment: Listening, Speaking, Reading, Writing and Comprehension. The raw score, scaled score, and proficiency level is listed for each of the five components.

Section D illustrates student performance in relation to the proficiency levels and to the State Average. The Proficiency Profile chart shows the scaled score "cut" line between proficiency levels Below Proficient (BP) and Proficient or Above (PA). Student ability is represented by the height of the solid bars and the striped bars show average ability statewide.

CONFIDENTIAL
SCHOOL ROSTER
English Language Proficiency (ELP) Assessment
Grade 4
A 2006 - 2007
ABC School

SYSTEM: ABC System (9999)

Test Date: Fall 2006

Student Name	Gender	Test Form	Listening (Max RS=22)			Speaking (Max RS=22)			Reading (Max RS: C1=15; C2=20)			Writing (Max RS: C1=15; C2=19)			Comprehension (Max RS: C1=31; C2=39)			Total (Max RS: C1=74; C2=83)		
			RS	SS	Prof	RS	SS	Prof	RS	SS	Prof	RS	SS	Prof	RS	SS	Prof	RS	SS	Proficiency Level
AYCOCK, JARON R. State ID#: 987600041 DOB: 01/03/1997	M	C2	16	101	PA	21	121	PA	6	84	BP	--	--	--	20	93	BP	43	383	Nearing Proficiency
BAY, MACEY V. State ID#: 987600040 DOB: 07/02/1997	F	C2	20	117	PA	22	136	PA	17	117	PA	17	137	PA	34	117	PA	76	449	Advanced
BOTELLO, BRENNEN C. State ID#: 987600039 DOB: 12/14/1995	M	C2	19	112	PA	22	136	PA	17	117	PA	16	127	PA	33	114	PA	74	441	Advanced
CALLIS, LARA L. State ID#: 987600038 DOB: 06/28/1997	F	C2	22	141	PA	21	121	PA	13	103	PA	12	104	PA	32	112	PA	68	423	Proficient
CASAREZ, SAGE R. State ID#: 987600037 DOB: 02/27/1997	M	C2	18	108	PA	22	136	PA	17	117	PA	17	137	PA	32	112	PA	74	441	Advanced
DRAIN, ARIELLE L. State ID#: 987600036 DOB: 10/15/1996	F	C2	18	108	PA	22	136	PA	15	109	PA	16	127	PA	30	108	PA	71	431	Advanced
FEE, DOMINIQUE P. State ID#: 987600035 DOB: 12/30/1996	M	C2	17	104	PA	22	136	PA	17	117	PA	17	137	PA	31	110	PA	73	437	Advanced
GARIBAY, FRANCES R. State ID#: 987600034 DOB: 04/25/1997	F	C2	20	117	PA	22	136	PA	13	103	PA	13	108	PA	30	108	PA	68	423	Proficient
GARRITY, TYREE M. State ID#: 987600033 DOB: 12/11/1996	M	C2	0	35	BP	--	--	--	--	--	--	--	--	--	0	33	BP	0	235	Novice
HELMICK, FELICIA J. State ID#: 987600032 DOB: 07/19/1997	F	C2	20	117	PA	21	121	PA	17	117	PA	13	108	PA	34	117	PA	71	431	Advanced
HENNESSY, KOBY L. State ID#: 987600031 DOB: 09/30/1996	M	C2	14	96	BP	14	93	BP	18	122	PA	10	96	BP	29	106	PA	56	401	Proficient

Legend: **RS**: Raw Score; **Max RS**: Maximum Possible Raw Score; **SS**: Scale Score; -- indicates test not taken
 Note: Any students who took the assessment with non-standard accommodations are marked with † symbol.

BP = Below Proficient PA = Proficient or Above

Page 1 of 1

The MontCAS ELP School Roster report lists all students—in a single school in a single grade—who took the MontCAS ELP in a certain year. The School Roster report includes the following information:

Section A shows the grade, the assessment year, the school name, and system name.


Section B lists each student alphabetically, along with his or her state student ID number, date of birth, and gender. The Test Form column identifies the specific test form administered to the students.

Section C lists each student's raw score (RS), scaled score (SS), and proficiency level (Prof), in each

language domain (Speaking, Listening, Reading, Writing, and Comprehension). Note that the Comprehension score is based on a subset of items from the Listening and Reading sections of the assessment. The language domain proficiency levels are: Below Proficient (BP) and Proficient or Above (PA).


Section D lists each student's Total MontCAS ELP raw score, total scaled score, and proficiency level: Novice (N), Nearing Proficiency (NP), Proficient (P), and Advanced (A).

Summary Report



MontCAS
(Montana Comprehensive Assessment System)
English Language
Proficiency Assessment

SYSTEM SUMMARY REPORT
English Language Proficiency (ELP) Assessment
Grade 4
2006 - 2007



Linda McCulloch, Superintendent
Montana Office of Public Instruction
P.O. Box 202001
Helena, Montana 59620-2001
www.opi.mt.gov

SYSTEM: **ABC System (9999)**

Test Form: **C1, C2**
Test Date: **Fall 2006**

Proficiency Level	Listening		Speaking		Reading		Writing		Comprehension		Proficiency Level	Total	
	Scaled Score Range	Number and Percent of Students	Scaled Score Range	Number and Percent of Students	Scaled Score Range	Number and Percent of Students	Scaled Score Range	Number and Percent of Students	Scaled Score Range	Number and Percent of Students		Scaled Score Range	Number of Students
Proficient or Above (PA)	At or Above 99	9 (90%)	At or Above 99	9 (100%)	At or Above 99	8 (89%)	At or Above 99	8 (100%)	At or Above 99	8 (80%)	Advanced (A)	At or Above 430	6 60%
Below Proficient (BP)	Below 99	1 (10%)	Below 99	0 (0%)	Below 99	1 (11%)	Below 99	0 (0%)	Below 99	2 (20%)	Proficient (P)	397 - 429	2 20%
											Nearing Proficiency (NP)	374 - 396	1 10%
											Novice (N)	Below 374	1 10%

N Students: 10 **N Students:** 9* **N Students:** 9* **N Students:** 8* **N Students:** 10

Mean Scaled Score: **Mean Scaled Score:** **Mean Scaled Score:** **Mean Scaled Score:** **Mean Scaled Score:**

System: 106.0 System: ** System: ** System: ** System: 102.4

State: 106.9 State: 110.6 State: 107.4 State: 110.5 State: 106.3

Median Scaled Score: **Median Scaled Score:** **Median Scaled Score:** **Median Scaled Score:** **Median Scaled Score:**

System: 110 System: ** System: ** System: ** System: 111

State: 108 State: 107 State: 109 State: 108 State: 106

N Students: 10†

Mean Scaled Score:

System: 409.4

State: 413.4

Median Scaled Score:

System: 431

State: 417

*Summary statistics exclude students who did not take this subtest.
** Less than 10 students
† Summary Statistics exclude students who took the assessment with non-standard accommodations.

Legend: **Mean Scaled Score:** The arithmetic average of a set of scaled scores. It is found by adding all the scores in the distribution and dividing by the total number of scores.

Median Scaled Score: The middle score in a distribution or set of ranked scaled scores. Half the scores in the set are below the median, and half are above it (the 50th percentile).

This report includes student information for less than 10 students and may not be distributed to the public under protection by The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) Use of the information by schools and teachers to assist students is encouraged.

The MontCAS ELP System and School Summary Reports show the distribution of scores by grade within a system or school. The reports are produced even if the number of LEP students in a particular grade is very small. Reports for less than 10 students include a footer indicating that they may not be distributed to the public; the student information is protected by The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99).

Section A shows the grade, the assessment year, and the system name.

Section B. For each language domain (Speaking, Listening, Reading, Writing, and Comprehension), the report shows—in the Number and Percent of Students columns—the number and percent of students whose scores placed them in each of the two Proficiency Level groupings: Below Proficient (BP) and Proficient or Above (PA).

Section C. The Total MontCAS ELP section shows scaled scores corresponding to each of 4 overall proficiency levels—Novice (N), Nearing Proficiency (NP), Proficient (P), and

Advanced (A). The Number of Students column shows the number of students whose performance placed them in each category and the Percent column represents that number as a percentage of the students in this grade who were tested. For example, the 2 in the Proficient (P) cell of the sample report above indicates that 2 students in the system scored in the Proficient (P) range, which is 20% of the students in this grade.

Section D. The N Students line shows the total number of students in the system in this grade for whom there is a language domain score and a total score. For example, the sample report shows that 10 4th-grade students took the Listening Test. The Mean Scaled Score line shows the average scaled score in each domain and overall for all tested students in the system. For example, the sample report shows that the mean scaled score on the Listening Test for this system was 106.0. The Median Scaled Score line shows the median scaled score in each domain and overall. The state mean and median are also shown for each domain and overall. Note that means and medians are shown only if N is 10 or greater.

Using MontCAS ELP Results

Monitoring Progress. MontCAS ELP test results can be used to determine whether students are making progress in developing English proficiency overall and within each language domain. To make comparisons between one year and the next, proficiency levels should be used. (Note that within a grade span, scaled scores can also be compared from year to year, as long as the student is being assessed with the same-letter form. Scaled scores cannot be used to monitor progress from year to year when students have moved to the next grade span, that is, in 1st grade, 3rd grade, 6th grade, and 9th grade.)

Informing Instruction. MontCAS ELP test results can be used to design instruction that capitalizes on students' strengths and addresses their weaknesses. Proficiency levels provide useful information on an individual student's profile across the language domains. For example, two students may both score as Proficient overall but have different strengths and weaknesses in the language domains. One may be lagging behind in Speaking, the other in Reading. With this information, instruction can be tailored to the individual student's needs.

